
Article:
Scott, J.G., Maini, P.K., Anderson, A.R.A. et al. (2019) Inferring tumour proliferative organisation from phylogenetic tree measures in a computational model. Systematic Biology. pp. 1-41. ISSN: 1063-5157

https://doi.org/10.1093/sysbio/syz070

Systematic Biology (2019), **0**, 0, *pp.* 1–41
doi:10.1093/sysbio/output

Inferring Tumour Proliferative Organisation from Phylogenetic Tree Measures in a Computational Model

Jacob G. Scott[1,2], Philip K. Maini[1,†], Alexander R.A. Anderson[3,†], and Alexander G. Fletcher[4,5,†,*]

[1] *Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Oxford, UK*
[2] *Departments of Translational Hematology and Oncology Research and Radiation Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio, USA*
[3] *Integrated Mathematical Oncology Department, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA*
[4] *School of Mathematics and Statistics, University of Sheffield, Sheffield, UK*
[5] *Bateson Centre, University of Sheffield, Sheffield, UK*
† – *contributed equally*

**Alexander G. Fletcher, email: a.g.fletcher@sheffield.ac.uk, telephone: +44 (0)114 2223846, address:*

School of Mathematics & Statistics, Hicks Building, Hounsfield Road, Sheffield, S3 7RH, UK

Abstract

We use a computational modelling approach to explore whether it is possible to infer a solid tumour's cellular proliferative hierarchy under the assumptions of the cancer stem cell hypothesis and neutral evolution. We focus on inferring the symmetric division probability for cancer stem cells, since this is believed to be a key driver of progression and therapeutic response. Motivated by the advent of multi-region sampling and resulting opportunities to infer tumour evolutionary history, we focus on a suite of statistical measures of the phylogenetic trees resulting from the tumour's evolution in different regions of parameter space and through time. We find strikingly different patterns in these measures for changing symmetric division probability which hinge on the inclusion of spatial constraints. These results give us a starting point to begin stratifying tumours by this biological parameter and also generate a number of actionable clinical and biological hypotheses including changes during therapy, and through tumour evolution.

Key words: Cancer, Evolution, Phylogenetics.

The cancer stem cell hypothesis (CSCH) posits that tumours are composed of a hierarchy of cells with varying proliferative capacities. Under this hypothesis, a subpopulation of 'cancer stem cells', also termed tumour initiating cells (TICs), are able to self-renew through symmetric division and also to differentiate into tumour cells resembling transit amplifying cells (TACs) through asymmetric division (Fig 1A), giving rise to the entire diversity of cells within a tumour (Fialkow et al. 1967). The CSCH provides a conceptual framework by which to understand many different aspects of cancer progression, including: the occurrence of functional heterogeneity despite genetically identical states (Magee et al. 2012, Sottoriva et al. 2010, Vlashi et al. 2011); resistance to chemotherapy (Chen et al. 2012, Werner et al. 2016) and radiotherapy (Bao et al. 2006, Dhawan et al. 2014, Diehn et al. 2009); recurrence (Dingli & Michor 2006); and metastasis (Pang et al. 2010). Despite its popularity, the CSCH has been the subject of continual debate and modification in order to maintain compatibility with experimental observations (Gilbertson & Graham 2012, O'Connor et al. 2014, Scott et al. 2019).

While the specifics of the CSCH are still a matter of debate, the clinical relevance of those cells with traits ascribed to TICs is clear. However, our ability to measure their dynamics in a clinical setting remains lacking. *In vivo* measurement efforts are limited to carefully conducted live imaging in genetically engineered mice (Ritsma et al. 2014), or genetic labelling and subsequent lineage tracing (Driessens et al. 2012). Although *in vitro* systems are better suited to the extraction of these parameters, to date little has been done to quantify them, as technically demanding single-cell lineage tracing (Lathia et al. 2011) is required. These experimental difficulties speak to the need for more theoretical work in this area, especially to propose metrics for quantifying proliferative parameters such as TIC symmetric division probability (Fig 1A) from clinical data. This is of particular importance as there is mounting evidence for the effect of proliferative hierarchy

on response to radiotherapy (Tamura et al. 2010) and chemotherapy (Chen et al. 2012), and microenvironmental factors such as hypoxia (Conley et al. 2012, Dhawan, Tonekaboni, Taube, Hu, Sphyris, Mani & Kohandel 2016), acidosis (Hjelmeland et al. 2011) growth factors (Doetsch et al. 2002), and even stromal cell co-operation or co-option (Liu et al. 2011, Vermeulen et al. 2010) have been shown to perturb this system. In summary, TIC symmetric division rate and somatic mutation rate are both specific parameters of interest in cancer biology, and these form the focus of the present modelling.

Several published mathematical models, using different formalisms and considering different aspects of heterogeneity, have predicted that the evolution of a solid tumour should depend strongly on whether or not it exhibits a proliferative hierarchy, and on the parameters of such a hierarchy. These models have included spatial proliferation constraints, microenvironmental heterogeneity and selective pressures, and the noted differences include shape, clonal heterogeneity, rate of evolution and growth dynamics. Werner et al. (2011) specifically studied the differences in bulk tumour behaviour between tumours arising from mutant TICs and TACs in a non-spatial context. In a spatial context, the work of Sottoriva et al. (2010), Sottoriva & Tavaré (2011), Enderling et al. (2009) and Morton et al. (2011) represent the first papers where it was shown that the parameters governing TAC dynamics can constrain tumour growth, and also to show that TIC-driven tumours have significantly different spatial growth patterns: specifically, that they exhibit 'patchy' growth. In none of these models, except Sprouffske et al. (2013), in which the main question centred on TAC numbers, were these differences studied across TIC symmetric division probabilities, which is a key parameter governing the hierarchy, and one that is exceedingly difficult to measure or perturb *in vitro* or *in vivo*.

To describe the evolutionary relationship between cells in a multicellular tissue, we require a phylogenetic approach. While the use of objective, genetic information to infer phylogenetic trees has a long history in evolutionary biology, its application to cancer evolution is much more recent, giving rise in the last decade to a subfield recently dubbed

'PhyloOncology' by Somarelli et al. (2016). Using phylogenies reconstructed from spatially separated biopsies and informatic algorithms, many aspects of tumour evolution have begun to be elucidated (Gerlinger et al. 2014), including the genetic heterogeneity present within a primary tumour (Sottoriva et al. 2013), the origin of individual metastatic tumours within the primary site (Gerlinger et al. 2012, Naxerova & Jain 2015), the earliest events driving progression and metastasis (Zhao et al. 2016), and the effect of chemotherapy on primary and metastatic sites (Faltas et al. 2016, Murugaesu et al. 2015).

In addition to these sorts of questions, there are precedents in other fields for using phylogenetic information, integrated with population dynamics to infer other underlying biological processes – a technique termed phylodynamics (Grenfell et al. 2004). For example, Leventhal et al. (2012) proposed that the phylogenetic tree contains a "fingerprint" that can be used to determine the evolutionary process driving the population in question. Modelling the spread of HIV within a contact network, the authors investigated whether the network structure could be inferred from the resulting disease phylogenies. To address this question, the authors simulated a range of epidemics on several families of random graphs and measured the resulting phylogenetic trees, finding that certain tree-based measures could discriminate between the qualitatively different families of random graph structures considered.

We may expect cancer cell phylogenies to look quite different to viral phylogenies. Nevertheless, these precedents motivate us to ask whether a similar approach could be used to discriminate between *in silico* tumours with different symmetric division rates. To test this hypothesis, here we study the effect of TIC symmetric division probability on tumour evolution using a computational modelling approach. We focus on observed patterns in reconstructed phylogenetic trees across a range of symmetric division probabilities. The estimation of this proliferative parameter from clinical data could help improve our understanding of the effect of therapies on tumour growth dynamics, and our ability to stratify tumours for consideration of different therapies. In this way, we seek to

provide translatable measures to aid in understanding tumour biology: to use mathematical modelling to 'see the invisible'.

The remainder of this paper is structured as follows. We first present a spatial stochastic model of tumour growth under a proliferative hierarchy with neutral mutations, which we embed on a two-dimensional lattice to enable the study of the effect of spatial constraints. Next, we develop an algorithm to reconstruct the branched phylogenetic structure from each realization of our tumour growth model. We apply a range of statistical measures of phylogenetic tree shape to simulation outputs for comparison. We explore the temporal dynamics of these measures over the course of tumour growth to assess whether they are robust to tumour size changes, and then to changes in mutation frequency. Finally, we discuss the possible clinical utility of these measures.

Materials and Methods

Model development

Here, we describe the development of a two-dimensional, lattice-embedded cellular automaton (CA) model of tumour growth with contact inhibition growing under neutral evolution and a proliferative hierarchy. We also develop a non-spatial companion model in order to assess the role of spatial constraints on the evolutionary process.

Proliferative hierarchy For both models, we consider a proliferative hierarchy comprising two cell types, TICs and TACs. We assume that each TIC divides symmetrically with probability α, creating two TICs, and asymmetrically with probability $1-\alpha$, creating one TIC and one TAC. For simplicity, we assume that α takes a constant value for all cells in a given simulation, and is not dependent on the mutation rate (see below). Note that in practice, microenvironmental parameters such as nutrient deprivation (Flavahan et al. 2013), acidity (Hjelmeland et al. 2011) and hypoxia (Heddleston et al. 2009, Li et al. 2009), as well as accumulated mutations such as

those commonly observed in colorectal cancers (Baker et al. 2014), are all known to be capable of affecting symmetric division probability among cells in a tumour. As it has been shown theoretically that the overall population dynamics of TIC-driven tumours is equivalent with or without TIC symmetric differentiation (Rodriguez-Brenes et al. 2011) (when a TIC divides to create two TACs), and as the lineage extinction possible in this case would significantly complicate our phylogenetic analysis, we make the simplifying assumption that there is no symmetric differentiation. We do not rule out that the addition of symmetric differentiation could affect phylodynamics, but leave that question for further study.

We assume that every TAC division is symmetric, creating two TACs, but only allow this to progress for β rounds of division, after which the TAC will die if chosen to divide again. Here β represents the replicative potential of TACs, and is posited to represent telomere length (Poleszczuk et al. 2014). Previous theoretical work has shown that tumour growth kinetics in spatially constrained geometries are strongly affected by the value of β (Morton et al. 2011). In particular, if $\beta > 5$, then simulated tumours experience unrealistically lengthy growth delays. Therefore we follow a previously used assumption (Sottoriva et al. 2010, Sprouffske et al. 2013) and fix $\beta = 4$. This mode of growth and differentiation is illustrated in Fig 1A. For simplicity, we neglect cell death, which could disrupt growth patterns. Indeed, Williams et al. (2016) have shown that the overall patterns of mutations, as measured by variant allele frequencies, is changed. The addition of cell death in this model is therefore a natural avenue for future work.

Neutral evolution To understand the effects of neutral evolution on tumours with differing proliferative hierarchies, we extend our model of tumour growth under a proliferative hierarchy to include random mutations. At each cell division, there is a possibility that one or more mutations occur. To determine the number of mutations accumulated by a given daughter cell, we independently draw a random number from a Poisson distribution with expectation λ. We assume for simplicity that every mutation

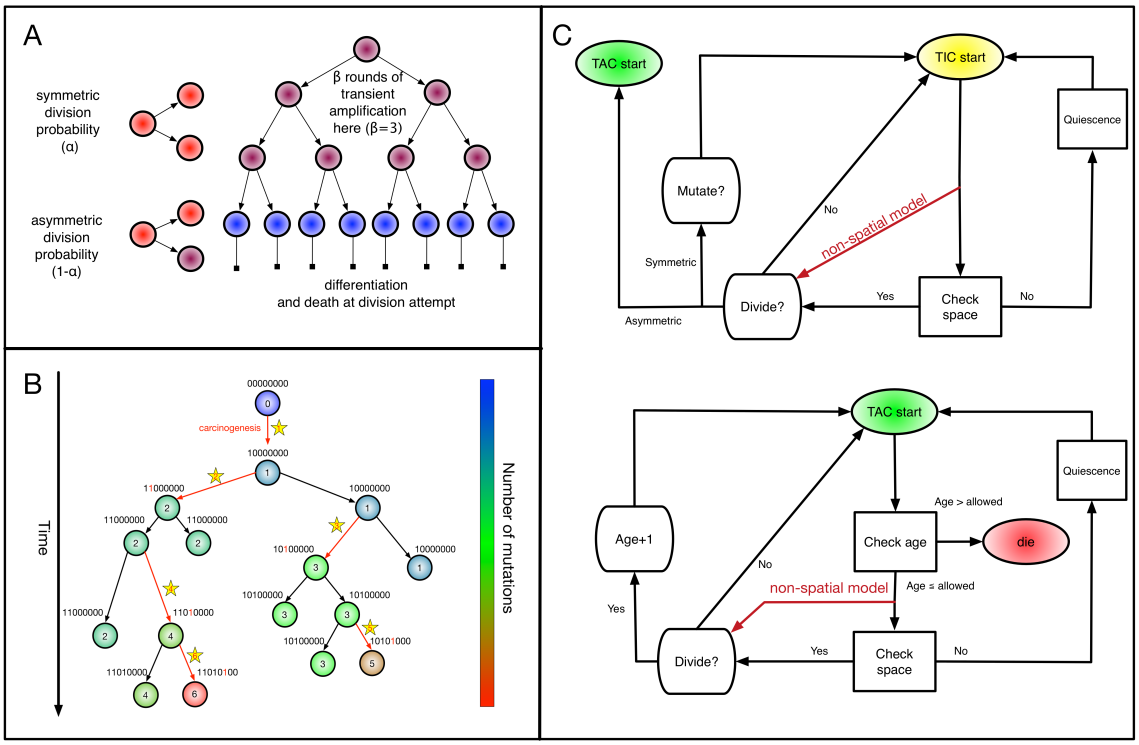

Fig. 1.

arising in our model is unique. This 'infinite sites' assumption is usually ascribed to Kimura (1969).

For simplicity, we assume that mutations confer no advantage, disadvantage or any other phenotypic change and therefore serve only as a method by which to track clonal lineages (i.e. they are neutral (Williams et al. 2016)). This assumption could in principle be loosened to allow for positive selection (Bignell et al. 2010), or a balance of positive and negative selection (McFarland et al. 2013). A schematic of this model of evolution, and labelling scheme, is shown in Fig 1B.

For computational efficiency, instead of storing a genome of length n ($\mathcal{G} \in \{0,1\}^n$), we record a unique integer flag only for the most recent mutation accumulated within a cell, which is passed down to its progeny, unless a mutation occurs, in which case a new flag (the next integer) is assigned. We also record each mutation event in the form of an ordered pair (parent flag, child flag), so that the complete genome, $\mathcal{G}$, can be

reconstructed for future use, and the length will be only as long as is needed to represent the changes which occurred during the simulation. As they are the only cells capable of forming tumours on their own, and infinite replication, we follow previous works in considering new mutations to accrue only in TICs (Sottoriva et al. 2010, Sottoriva & Tavaré 2011, Sprouffske et al. 2013, Poleszczuk et al. 2015). A natural extension for human cancer would be to consider instead sequences formed of DNA characters ATCG.

Non-spatial model implementation

To implement a first version of our hierarchical model with neutral evolution, we consider the case in which there are no spatial constraints – which could be a sensible model for liquid tumours like leukemias which are well mixed in the blood. We initialize our simulation with a single TIC. We then implement discrete updates, which we will term 'time steps'. As we are not studying temporal dynamics here, we do not prescribe any proper dimension to this time. At each discrete time step we choose a cell uniformly at random from the population to either divide and possibly mutate, or die. Our simulation is considered complete when the total population of cells reaches a prescribed number.

If the cell is a TIC, then we first draw a random number, r, from $\mathrm{U}[0,1]$ and compare it to the probability of symmetric division, α. If $r < \alpha$, then the cell divides symmetrically, and we draw the number of new mutations accumulated in each daughter independently from Poisson(λ). A new TIC daughter is then generated and given a mutational 'identity' which is the mathematical sum of the parent cell's identity and the number of new mutations. The parental cell is also updated by changing its identity based on the number of new mutations it accrues, if any. If the cell is not determined to divide symmetrically, then it divides asymmetrically and a TAC daughter is created with age 0 and the parental TIC is updated for new mutations as above, see Fig 1.

If the chosen cell is instead a TAC, then we first check that its age is less than the allowed TAC age, β. If the age is equal to the allowed TAC age, then the cell dies. If the age

is less than the allowed TAC age, then the cell ages and divides, and a new daughter with the same identification is created, whereupon the loop continues (see Fig 1 for schematic).

Spatial model implementation

As we are interested in the effect of the proliferative hierarchy on the neutral evolutionary process in solid, spatially constrained tumours, we embed our cell-based model in a two-dimensional square lattice. While recent work has shown some qualitative differences in vascularised CA models between two and three dimensions, using a two-dimensional lattice for unvascularised tissue is a common simplification (Anderson & Chaplain 1998, Alarcón et al. 2006, Gerlee & Anderson 2008, Scott et al. 2016) that allows spatial constraints to be studied in a computationally tractable manner. In addition to the above description of cell proliferation, we consider cell proliferation to be modulated by contact inhibition (Anderson 2005). A cell is allowed to divide only if at least one of the eight adjacent lattice sites (north, south, east, west, and diagonal neighbours) is unoccupied; if this is not the case, then we consider the cell to be in a quiescent state that may be exited when space becomes available. At each time step, each 'cell' has an opportunity to divide given that it has space to do so. Cells are chosen uniformly at random for updates from the entire population to avoid order bias. Apart from these spatial effects, the model is otherwise identical to the non-spatial model presented earlier.

Cell-type specific rules If space is available, and the cell is a TIC, then the type of division is determined by choosing a uniform random number, r, from $[0, 1]$. If $r < \alpha$, then the TIC divides symmetrically, creating another TIC that is placed uniformly at random in one of the free neighbouring lattice sites. The parent and daughter TICs will independently acquire a random number of new mutations, as described above. If $r \geqslant \alpha$, then the TIC divides asymmetrically, creating a TAC that is placed uniformly at random in one of the free neighbouring lattice sites. The daughter TAC is created with the same mutation

identity (ID) as the parent, and age = 0, while the parent TIC will independently acquire a random number of new mutations, as described above.

If the chosen cell is instead a TAC, then the check after available space is a check of the cell's proliferative age, which is the number of divisions as a TAC. If the TAC age is equal to the replicative potential, β, then the TAC dies, at which point it is removed from the simulation. If the TAC age is less than β, then we create a new TAC daughter and place it uniformly at random in one of the free neighbouring lattice sites. The parent and daughter TACs share the same mutation ID and their age is updated to be one more than the age of the originally chosen TAC.

Full implementation

The full CA flow-chart, represented in Fig 1C, schematises the entire process of cell fate decisions that each cell undergoes at each time step in the spatial model. In the top panel, the rule set followed by the TICs is represented to include differentiation and

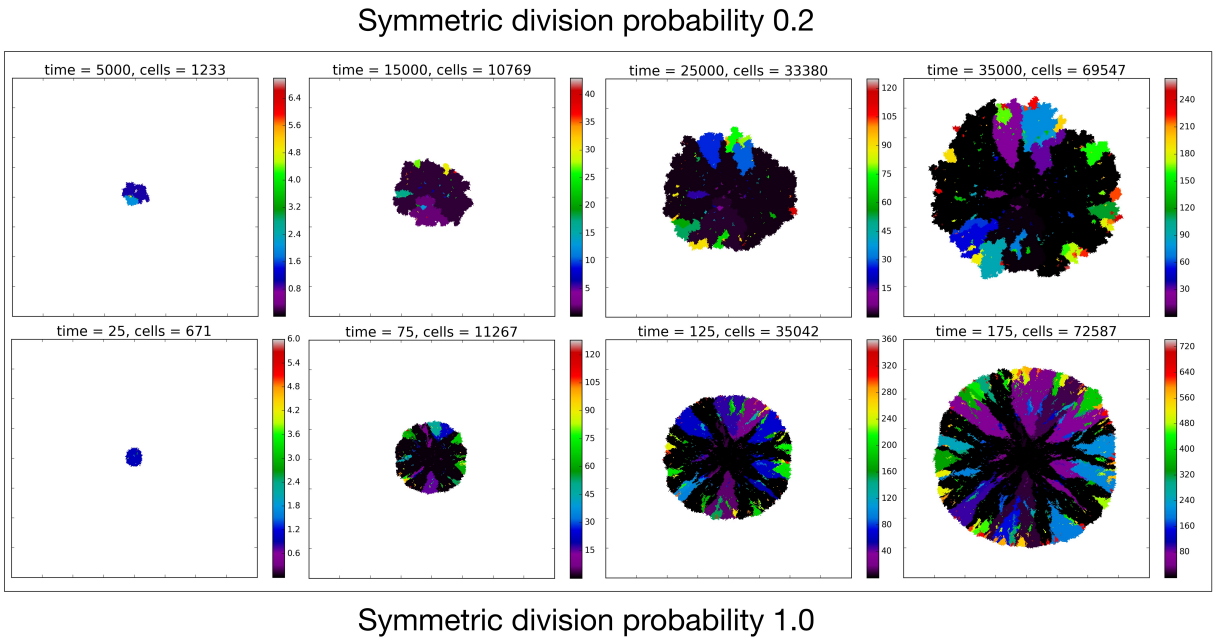

Fig. 2.

mutation. In the bottom panel, the TAC rule set is defined to include death by terminal differentiation and TAC aging. An example simulation of tumour growth over time is shown in Fig 2, where the effect of lowering α can be seen on overall tumour growth kinetics, where the colour-bar represents the current clonal state (mutation ID) of a given clone.

Recovering phylogenetic trees from simulation

While experimentalists and clinicians can only infer phylogenies from incomplete data, reconstruction of the 'true' phylogeny is possible in our model as we can record the entire life history of the simulated tumour. Thus, we can test whether phylogenetic tree-based measures are able to discriminate TIC symmetric division probability in the case where the 'ground truth' is known. At each time step we record the spatial location of each individual cell with its mutation ID, which is our CA state vector. Additionally, we record the evolutionary 'life history' as a list of ordered pairs of every mutation event (parent mutational ID, child mutational ID). We then recursively construct the phylogenetic tree from this life history.

Phylogenetic tree reconstruction algorithm To create the complete tree data structure required for our quantitative analyses we use the information encoding the mutation events from our stochastic simulation. To this end, we create a list of unique parent-child pairs using the life history of mutation events. We then apply an iterative process in which each child is added as a subnode below the parent (from the unique parent-child pair). This process is continued until all parent-child pairs are added to the structure, and the tree is complete. The simulation code and functions to create these trees and calculate the metrics is freely available on request.

Qualitative comparison of reconstructed trees To compare phylogenies from simulations with different underlying parameter values, we first construct and visualize the

phylogenies constructed from three example simulations with differing TIC symmetric division probabilities in Fig 3. It is clear by inspection that the number of mutations increases with symmetric division probability (more branches). However, the tree structure is not as easy to parse visually. For ease of visualization the trees depicted in Fig 3 have been pruned of all terminal nodes (also called leaves) with no children of their own. While this transformation does affect the quantitative results, it does not qualitatively affect the resultant phylogenetic tree statistic ranks (see Fig 8). All analyses shown will utilize the full trees.

Candidate tree-based measures for model comparison

Visual inspection of Fig 3 suggests that simulations with different TIC symmetric division probabilities generate distinct phylogenetic trees. However, to draw meaningful conclusions we must perform a quantitative comparison. Here we present several measures useful in summarising and comparing phylogenetic trees. The most commonly studied property of a phylogenetic tree's shape is its balance, defined as the degree to which internal nodes (branch points) have the same number of children as one another. Balance (or imbalance) indices depend only on the branching topology of trees, and not on other factors like branch length or other features of the terminal branches (leaves). Since the first balance index by Sackin (1972), many others have been proposed with slightly differing properties (Mir et al. 2013). One of the first papers to present a systematic comparison of a suite of balance indices (often denoted with the letter 'B') and indices of imbalance (denoted with 'I') was by Shao & Sokal (1990), who reported striking differences between the studies' measures. Their central message was that different measures on trees can give insight into different aspects of the underlying processes governing the interactions, and one should thus consider several measures for any given tree or family of trees. In this study we will consider several tree topology-based measures.

Before describing the measures, it is worthwhile to briefly define the terms which

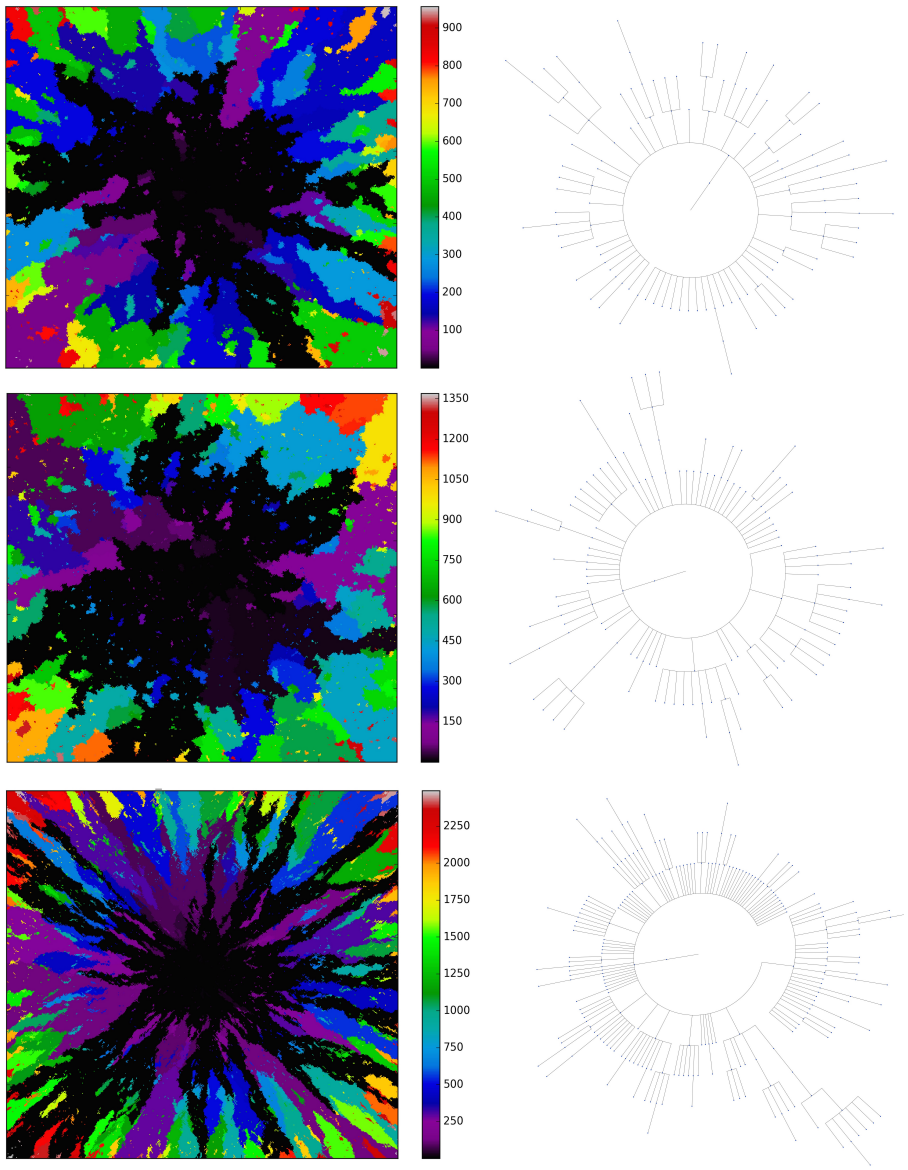

Fig. 3.

are used to describe trees, and the two basic underlying stochastic models which have been proposed to describe neutral evolution and the resulting topologies. Phylogenetic trees describe the evolutionary relationship between individuals with different traits from one another, or in the case of our model, different mutational combinations (genotypes). In our model, each simulation begins with a cell with mutation flag 1, or a genotype with the first allele mutated (1000...), termed the 'root', and evolution progresses stochastically, by

adding individual mutations at subsequent alleles and increasing the mutation flag, as described in Fig 1B. At each mutation event, an evolutionary branch point is created, which is termed a node in phylogenetic tree terminology. If this node gives rise to no other children during the simulation, it is termed a terminal node, or leaf. There are two common, classically referenced models, which bear mention here as well, since many tree topology-measures are normalized against them. The first, described by Yule (1925) and sometimes termed the 'equal rate Markov' model, begins with a single root and proceeds by replacing, uniformly at random, a given leaf with a node with two children of its own. The process continues until the desired number of leaves exist. The other main model, termed the 'Proportional to Distinguishable Arrangements' or uniform model, was described by Rosen (1978). This model, which is truly a model of tree growth rather than an explicitly evolutionary process, begins as does the Yule model (and indeed ours) with a single node labelled 1. At each update step, a new leaf is added to the tree at any point, either internal node or leaf. These models will serve as normalisation factors in several of the measures we present below, which are summarised graphically in Fig 4.

Sackin index The Sackin index was the first statistic used to understand the balance of a phylogenetic tree (Sackin 1972, Shao & Sokal 1990). To compute this statistic, one sums the number of ancestors (N_i) for each of the n terminal nodes of the tree:

$$I_s^n = \sum_{i=1}^{n} N_i. \tag{0.1}$$

This index increases with tree size: under the Yule growth model, its expectation $\mathrm{E}[I_s^n]$ grows as $2n \log n$ (Yule 1925). One can therefore only perform a meaningful comparison of Sackin indices of trees generated from tumours if they are the same size.

Normalized Sackin index To address this dependence on tree size, several normalisations to the Sackin index have been proposed, two of which we explore here. In particular, one can normalise the Sackin index of a phylogenetic tree to the expectation

value of a similarly sized tree, under the Yule growth model:

$$I_{Yule} = \frac{1}{n}\left(I_s^n - 2n\sum_{j=2}^{n+1}\frac{1}{j}\right). \tag{0.2}$$

One can alternatively normalise using the Proportional to Distinguishable Arrangements (PDA) model (Aldous 1996, 2001, Rosen 1978) which is simply the Sackin index scaled by $n^{3/2}$.

The B1 statistic The B1 statistic, originally described by Shao & Sokal (1990), considers the balance of a tree. To calculate the measure, one uses all i internal nodes of the tree with the exception of the root (the founding cell). For each non-root internal node j, the maximum number of nodes traversed along the longest possible path to a terminal node, M_j, is counted. The B1 statistic is then defined as

$$B1 = \sum_i \frac{1}{M_j} \qquad \forall i \neq root. \tag{0.3}$$

$\bar{N}$ The $\bar{N}$ statistic reports the average number of nodes above a terminal node. To compute this, we sum the path from each terminal node to the root, and divide by the number of terminal nodes. An alternative definition is the Sackin index 'normalised' by the number of terminal nodes. We define this as

$$\bar{N} = \frac{1}{n}\sum_{i=1}^{n} N_i, \tag{0.4}$$

where n is the number of tips and N_i is the number of internal nodes between tip i and the root of the tree. For a more complete review and comparison of the measures presented here, and others, see Blum & François (2005) and Shao & Sokal (1990).

Examples of how these measures change on several example trees with equal numbers of leaves (but different numbers of internal nodes) are presented in Fig 4. (Note that in contrast with these polytomous trees, our tumour growth model does not exhibit polytomies, since all cell divisions are dichotomous and each node in the phylogenetic tree is defined by a unique mutation in an infinite sites model.) In these examples, we compute

each of the presented measures for comparison. From left to right, the trees contain 4, 3 and 2 internal nodes, respectively, but the same number (6) of leaves. We note that the measures do not all follow the same pattern. For an exhaustive description of all possible trees with 6 leaves, and the correlation of a larger family of associated measures, see Shao & Sokal (1990).

Results

Measuring trees from simulation

As our primary goal is to identify whether tree-based measures allow discrimination of simulated tumours with different TIC symmetric division probabilities, we focus on changes in tree measures as we vary comparable simulations changing only this parameter. To compare the model tree measures, we first perform 50 stochastic simulations of both our non-spatial and spatial CA using a range of TIC symmetric division probabilities ($0.2, 0.4, 0.6, 0.8$ and 1.0), holding mutation rate and TAC lifetime constant ($\lambda = 0.01$ and $\beta = 4$). For each simulation, we construct the resulting phylogenetic tree at tumour size $250,000$ cells, as described in the Materials and Methods section. We then measure the

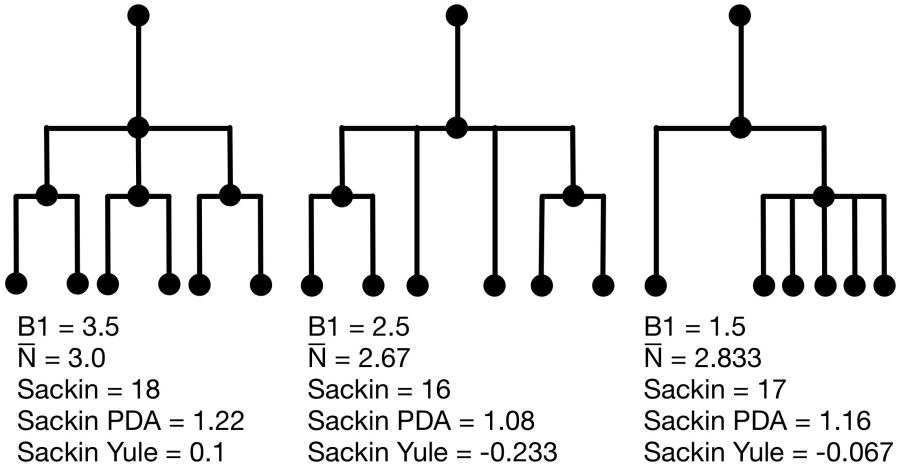

Fig. 4.

value of each summary index defined earlier for all 50 simulations at the final time point and plot the distribution in a box-whisker plot, which is shown in Fig 5 with each data point overlaid in a swarm. Differences between distributions were determined using the Wilcoxon rank sum test. While these statistics were performed *post hoc*, we should note that standard statistics can be misleading for simulation based studies with arbitrarily large sample sizes (White et al. 2014) (see Supplementary Fig 9 for effect size).

Variation of tree-based measures with symmetric division probability

The results of the model are presented in Fig 5. We find that all of the indices have monotone relationships with symmetric division probabilities except for $\bar{N}$ in the spatial model, and B1 in the non-spatial model. Of note also, is that only in the Sackin index do we see qualitative agreement for the spatial and non-spatial models (monotone up/down) for both in the standard (normalized) Sackin model. This difference in utility of the different models for spatial and non-spatial is not unexpected, as Shao & Sokal (1990) have previously shown that even for similar questions, different models of tree topology will have different uses. As our primary purpose is to understand the spatial cancer model, we will leave a deeper investigation into the dynamics of the non-spatial model for future work, and concentrate our analysis from here forward on the spatial model.

In terms of discernibility for the spatial model, of the normalised indices the B1 statistic has the least overlap in error between symmetric division probabilities (i.e. comparing the cases $\alpha < 1$ with the case $\alpha = 1$). All measure distributions are significantly different by the Wilcoxon rank sum test ($p < 0.05$) except 0.4 and 0.6 in the Sackin index normalised by the Yule model ($p = 0.08$). While we recognize the dangers in reporting p-values in simulation based studies (White et al. 2014), we report them here for comparison, and report effect size as well, with full statistics for both the spatial and non-spatial model reported in Fig 9. The strongest effect for the spatial model is seen in the Sackin index ($R^2 = 0.871$), followed by the Yule normalised Sackin index ($R^2 = 0.743$).

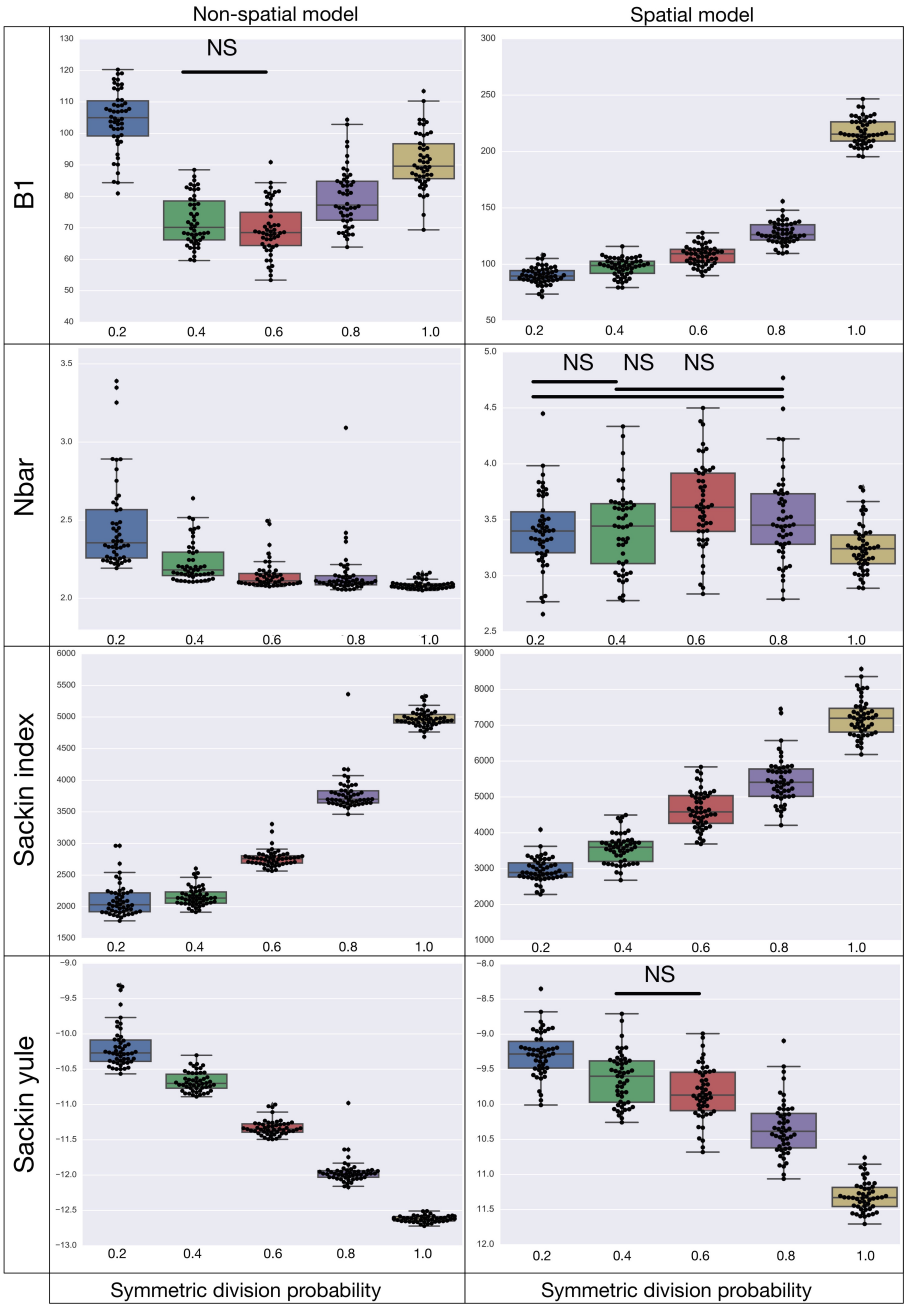

Fig. 5.

Dynamics of tree-based measures during tumour growth

As discussed in Materials and Methods, the measures considered here are strongly dependent on the total number of nodes in the tree. With all other parameters held

constant, simply allowing a tumour to grow larger would increase the number of total mutations, and therefore the number of total nodes, subsequently altering the value of the measure. To ensure that the differences we have noted are robust to changing tumour size, we next consider how these measures evolve during the growth of a tumour.

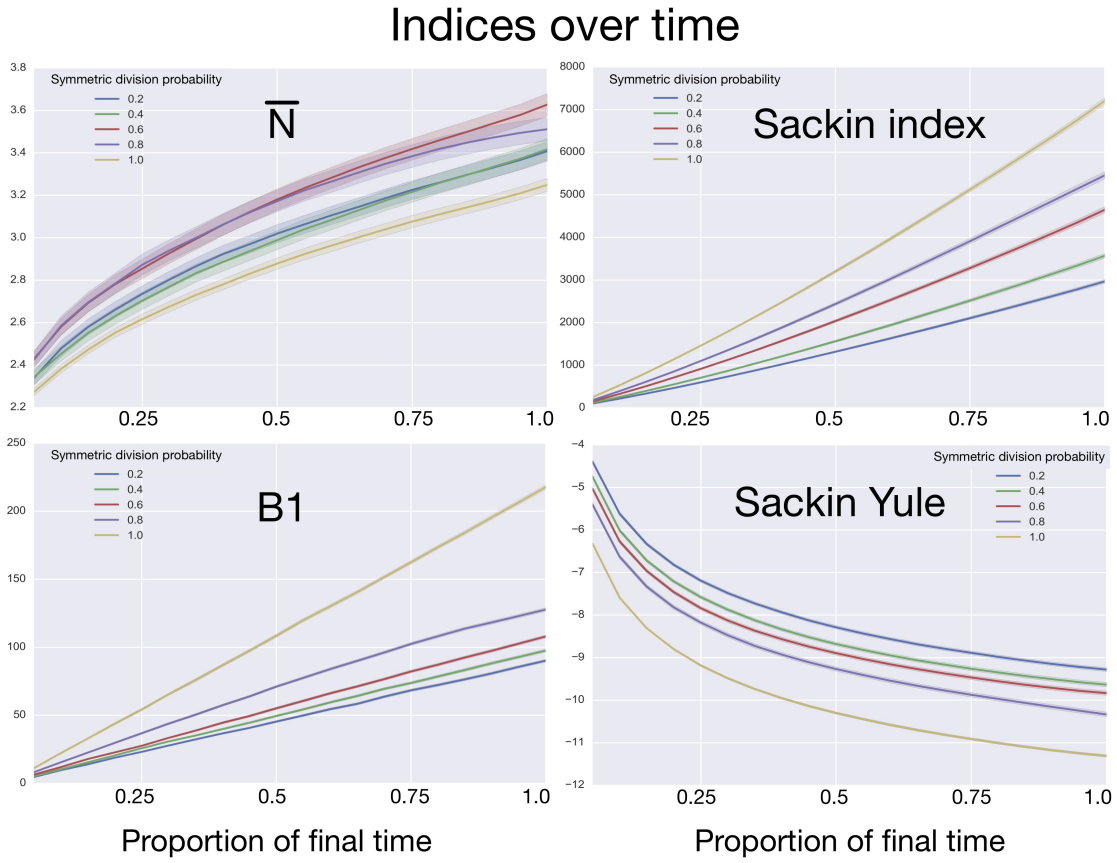

Fig. 6.

To determine how these measures vary over the lifetime of a growing tumour, we measure each index over the course of multiple simulations. To accomplish this, we use the life history to reconstruct the phylogenetic tree at 20 equally spaced time points during the course of 50 simulations for each value of the symmetric division probability. Note that since the time taken for each simulated tumour to fill the spatial domain depends strongly on the symmetric division probability (see Fig 2), to compare 'like for like' we break each

life history into equally spaced time intervals, which act as surrogates for tumour size. Comparing across tumour size is of greater utility, clinically, since while the age of a given tumour is rarely known, size can be readily estimated.

After reconstruction, we then create a 'time' trace for each statistic. We plot these statistics over 'time' in Fig 6, where each family of 50 simulations (for a given symmetric division probability) is represented by a single trace with the standard deviation represented by the coloured error bars. We find that for each of the statistics, except $\bar{N}$, the relationships between the symmetric division probabilities are maintained over time, suggesting that, if we know the tumour size, and true phylogeny, we can estimate the relative symmetric division probability between two samples from these measures. This statement must be somewhat qualified by the fact that mutation probability was also held constant for these simulations. While estimating mutation probability is not trivial, significant advances have been made in measuring the speed of the 'evolutionary clock' of tumours: essentially a proxy for mutation probability (Curtius et al. 2016). Further, we found that the rank order of each discriminatory measure holds throughout tumour growth, indeed becoming more discriminatory as the tumours grow larger (with the exception of $\bar{N}$). As the tumours simulated in this study are unrealistically small given the computational constraints, this information gives us hope that in tumours of realistic size, these measures would be even more useful. This becomes particularly important as the statistics that we have calculated come from the 'true trees', that is, trees comprised of all mutation events. In reality, trees would be inferred from the imperfect information gleaned from biopsies.

Dependence of tree-based measures on mutation probability

As the tree measures depend heavily on the number of mutations within a given tumour, and therefore the number of branches within a given tree, we next ask how these measures behave when we vary mutation probability (λ) and symmetric division

probability simultaneously. To answer this, we perform 10 stochastic simulations for each combination of the symmetric division probabilites considered previously and 5 different values for λ varying over two orders of magnitude $(0.001, 0.005, 0.01, 0.05, 0.1)$, which spans most of the range (per cell per division) as noted by Alexandrov et al. (2013). We then use the previously described method to reconstruct the resulting phylogenies and calculate the measures previously discussed. In particular, we ask how the Sackin index, the B1 statistic and the normalized Sackin index perform over this range of λ to better understand the applicability of these measures in determining differences in symmetric division probability.

We plot the results of this parameter investigation in Fig 7. In each heat map, we plot the mean of the 10 simulations for each parameter combination with symmetric division probability varied along the horizontal axis and mutation probability along the vertical axis. The indices which are not normalized by branch number, namely the Sackin index and B1 statistic, increase monotonically with mutation probability and symmetric division probability in all cases. The Sackin index normalised by the PDA model, however, varies somewhat unexpectedly and has a global minimum at symmetric division probability of 1.0 and mutation probability 0.01. This measure is monotonic in symmetric division probability except at the highest mutation probability where it becomes somewhat more difficult to determine the differences. As before, the B1 statistic appears to be the most stable, and only breaks down slightly in its ability to distinguish between the families of simulations at the lowest mutation probability ($\lambda = 0.001$) and the middle range of symmetric division probability (symmetric division probabilities $= 0.4 - 0.8$), as can be seen in Fig 7. In these ranges of the parameter space our model may not provide useful predictive power.

Discussion

While the use of phylogenetic trees is increasing in translational oncology laboratories, there has yet to be a method found by which we can utilise the information

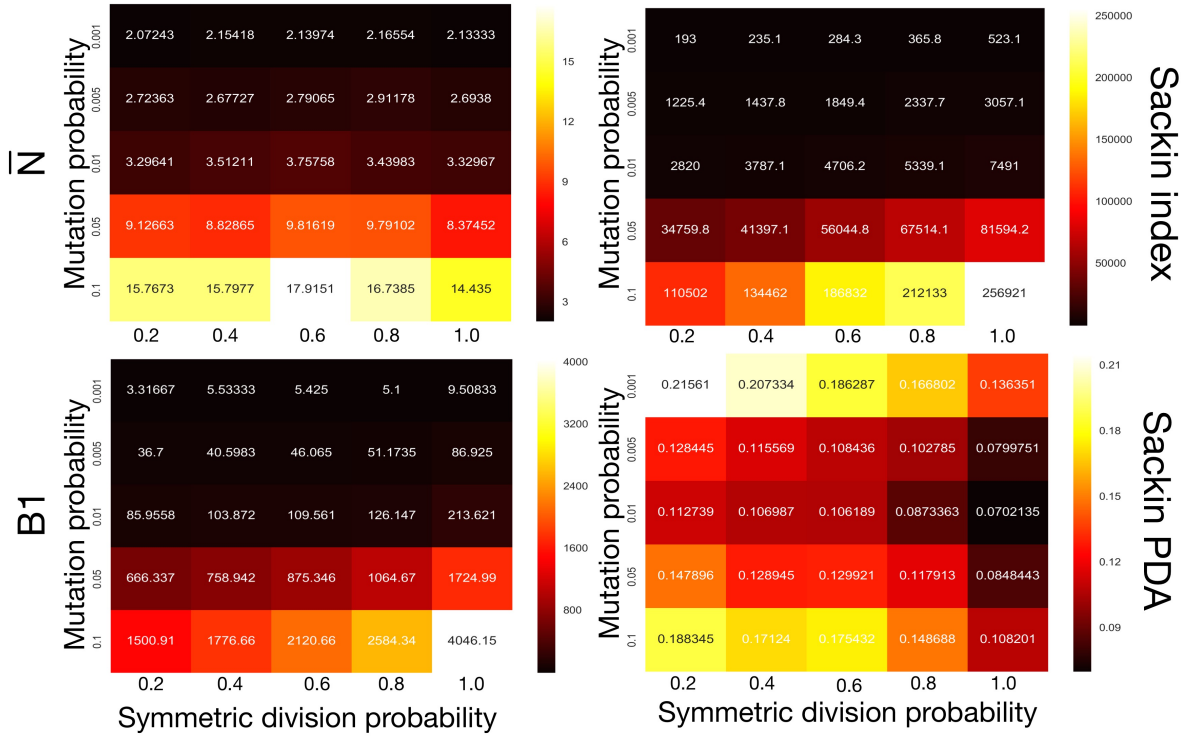

Fig. 7.

clinically. To address this shortcoming, we worked to leverage the growing interest in biomarker derivation from spatially distinct tumour biopsies (Dhawan, Graham & Fletcher 2016), and the recent success of Leventhal et al. (2012) and others in teasing apart complex biological rules from phylogenetic information. We developed an individual based model of tumour growth under a TIC driven proliferative heterogeneity which undergoes neutral evolution. We then developed an algorithm to construct phylogenetic trees from simulated tumours. The resultant trees were then analysed and compared using a suite of statistical measures of tree (im)balance. Through this method, we have generated a large dataset that includes the observed statistical measures of the 'true' phylogeny for tumours spanning the full range possible of symmetric division probabilities, which we feel is appropriate as symmetric division probabilities as low as 5% have been reported in glioblastoma (Lathia et al. 2011) and as high as 90% have been reported in colorectal cancer (Baker et al. 2014).

It is worth noting that, as our phylogenies are exact, some of the conclusions one may draw in the future from empirically inferred trees would have to be tempered by the inherent uncertainty in the inference process. Further, we are by no means the first group to seek to utilize simulation to study tree topology, with notable recent examples being INDELible, a flexible platform to simulation insertions and deletions (Fletcher & Yang 2009), as well earlier models of simulating neutral Wright-Fisher models (Hudson 2002).

In particular, we compared the classical measures of tree topology – the Sackin index and the B1 statistic – as well as normalized versions of each across several parameters of our spatial and non-spatial models as well as through the process of tumour growth. Not surprisingly, we found that the Sackin index was able to discriminate between the families of simulations as it is directly correlated with branch number (in this case correlating with total number of mutations in the TICs, which also is increased with increasing symmetric division probability). Encouragingly, we also found that the normalised version of this metric was able to discriminate between the different symmetric division probabilities, suggesting a more meaningful (and measurable) topological difference between the underlying phylogenetic trees resulting from these parameter changes (representing diverse biological traits).

While we have shown that these measures differ significantly from one another, we have not yet provided a method by which we can use the metric of a given tree to directly predict the symmetric division probability of an unknown tumour. However, the present work at least allows us to understand the rank order of symmetric division rate for two tumours given their measured indices. This could be particularly useful in certain clinical settings. For example, this could allow us to determine how a given therapy affects symmetric division probability by using our calculated measures over serial biopsies, and subsequent phylogenetic reconstruction. This could prove particularly useful in the treatment of leukemias, where the target cell is known to be the TIC, eradication of which is a requirement for cure (Roeder et al. 2006). In this case, after phylogenetic

reconstruction using any of several methods available for single sample whole genome sequencing (Carter et al. 2012, Roth et al. 2014, Deshwar et al. 2015), the tree topologies could be compared before and after therapy, giving a measure of relative change. Our metric (derived in this case from our non-spatial model results) could therefore prove a useful adjunct to existing methods of predicting TIC fraction (Werner et al. 2016) to determine therapeutic efficacy, and guide therapy breaks or switching.

Even with state-of-the-art multi-region sequencing approaches, most reconstructed cancer phylogenies are relatively small (Gerlinger et al. 2014, Zhao et al. 2016) with very few leaves, preventing the application of our statistical method in its current form. With continued advances in single-cell sequencing, and more examples of higher spatial sampling from larger tumors, like in the most recent TRACERx Renal study (Turajlic et al. 2018), the situation may change, but it is worth reiterating that this theory has not yet been shown to be quantitatively accurate in real tumor samples. Even in the case that it does not become effective in the near future, however, we assert that beginning to use metrics of tree topology to compare tumors before and after therapy, or across grade or survival, could prove useful to enhance our understanding of tumour evolution and treatment response in the near term.

Aiming towards a translatable method by which to infer the symmetric division probability in solid tumours, we have identified several phylogenetic tree based measures that correlate with TIC symmetric division probability. We have found several measures which are able to discern differences in simulated tumours between symmetric division probabilities. These results are robust to changes in tumour size, specifically maintaining their rank throughout tumour growth. The rate of mutation does affect these results to some degree, but rank is maintained permitting comparison through time, or between tumours of similar size.

While there is some overlap amongst the measures when more than one parameter is varied, with information on mutation probability and tumour size, relative symmetric

division probability can be estimated. We have restricted our focus to measures of (im)balance, a basic property of phylogenetic trees based only on their branching topology. With more information, such as evolutionary branch lengths (Kirkpatrick & Slatkin 1993, Mooers & Heard 1997) which are linked to the 'speed' of a tumour's molecular clock (Curtius et al. 2016), some of these limitations could be obviated. Further, we have only considered neutral evolution. While most tumour evolution is likely neutral (Williams et al. 2016), there is certainly evidence for non-neutrality in the form of driver and passenger mutations (McFarland et al. 2013, 2017), which would drastically affect the resulting phylogenetic trees (Grenfell et al. 2004) – especially with intervening treatment regimens. How non-neutral evolution and treatment affect our measures remain avenues for future work.

Acknowledgements

The authors thank Trevor Graham and Helen Byrne for insightful comments and discussions. AGF is supported by a Vice-Chancellor's Fellowship from the University of Sheffield.

Supplementary Material

Pruning trees does not affect rank of statistics

To visualize the trees more easily in Fig 3, we prune the leaves from each full tree. While this changes the absolute value of each of the tree-based measures, it does not affect their relative ranking. This suggests that each measure is capturing something fundamental about the biology as it appears invariant with tree size. This is corroborated by the results shown in Fig 6, indicating that the rank of each measure is stable over tumour growth.

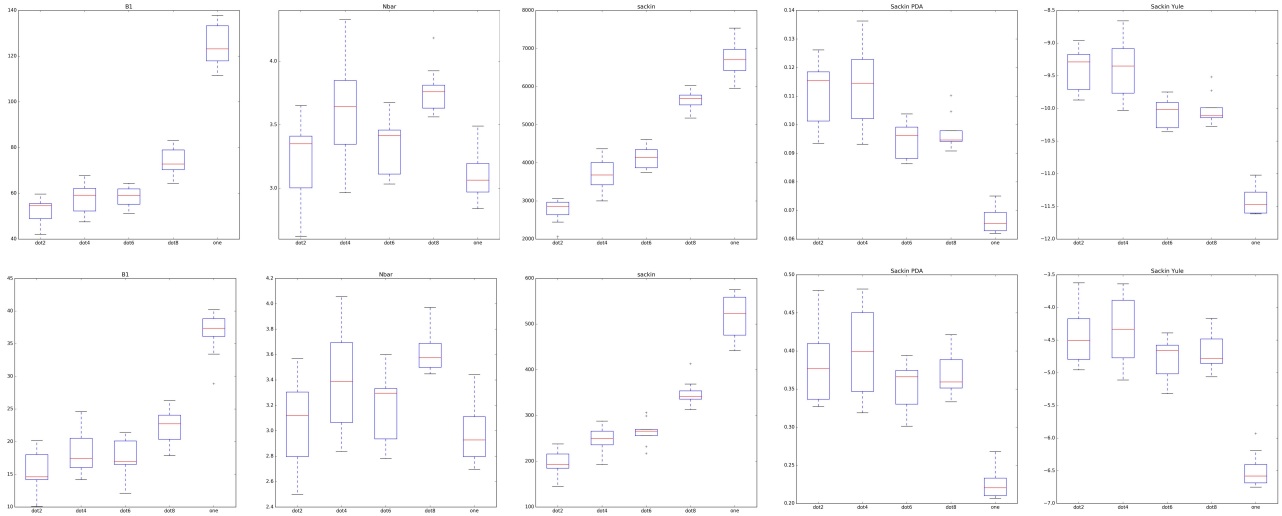

Fig. 8.

Effect size of symmetric division probability

To better understand the impact of the symmetric division probability on changes in the resulting tree topology, rather than just use differences between families of simulations, we compute the regression slope, R^2 and p-value of the regression line for each case. For the B1 statistic we find a regression slope of 142.64, $R^2 = 0.72$, $p = 1.74 \times 10^{-71}$, and in the non-spatial model a regression slope of -13.2, $R^2 = 0.056$, $p = 1.7 \times 10^{-5}$. For the Sackin index we find a regression slope of 5178.61, $R^2 = 0.871$, $p \approx 0$, and in the non-spatial model a regression slope of 3690.92, $R^2 = 0.8673$, $p \approx 0$. For the Yule normalised Sackin index we find a regression slope of -2.380, $R^2 = 0.743$, $p = 3.25 \times 10^{-75}$, and in the non-spatial model a regression slope of -3.118, $R^2 = 0.948$, $p \approx 0$. For the $\bar{N}$ statistic we find a regression slope of -0.111, $R^2 = 0.0075$, $p = 0.172$, and in the non-spatial model a regression slope of -0.457, $R^2 = 0.303$, $p = 3.33 \times 10^{-21}$. These values are plotted in Fig 9.

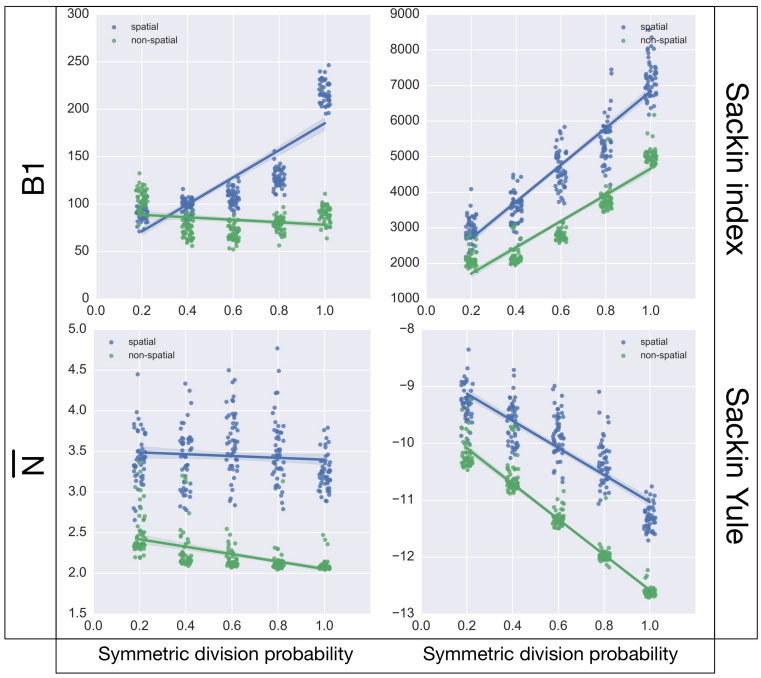

Fig. 9.

Algorithm for generating individual cell 'genomes' from mutational flag and life history

Our algorithm to reconstruct individual cell 'genomes' from the mutational flag and life history allows for significant increase in speed of our tumour growth model and reduced memory requirements by several orders of magnitude. While there are many tree reconstruction algorithms available to infer phylogenies from real data, most notably PHYLIP (Felsenstein 1981), as our data are 'perfect data' from our simulation, and therefore in a bespoke data structure, there is no inference required, and the task is a simple recursive tree building algorithm, which proceeds per the pseudocode below. Code is available at: `https://github.com/cancerconnector/clonal_evolution.git`

Algorithm 1: Pseudo-code describing algorithm to reconstruct genomes from unique mutation flags and family history.

```
Data: Dictionary of unique Parent:Child pairs and spatial array of unique
      mutation flags at time point of interest.
Result: Array of bitstrings representing 'genomes' of cells in array.
for All cells in array do
    if mutation ID = 0 then
        break
    end
    set bitstring to '1' + maxval(mutation ID) '0';
    final-parent = 2;
    if mutation ID = 1 then
        finalize bitstring
    end
    while final-parent > 1 do
        final-parent = lookup parent(cell of interest) in dictionary;
        flip bitstring at position(cell of interest) to '1';
    end
    finalize bitstring;
end
```

References

Alarcón, T., Owen, M., Byrne, H. & Maini, P. (2006), 'Multiscale modelling of tumour growth and therapy: the influence of vessel normalisation on chemotherapy', *Comp Math Methods Med* **7**(2-3), 85–119.

Aldous, D. (1996), Probability distributions on cladograms, *in* 'Random Discrete Structures', Springer, New York, NY, pp. 1–18.

Aldous, D. (2001), 'Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today', *Statist Sci* **16**(1), 23–34.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Aparicio, S., Behjati, S., Biankin, A., Bignell, G., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A., Caldas, C., Davies, H., Desmedt, C., Eils, R., Eyfjörd, J., Foekens, J., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D., Jones, D., Knappskog, S., Kool, M., Lakhani, S., López-Otn, C., Martin, S., Munshi, N., Nakamura, H., Northcott, P., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J., Puente, X., Raine, K., Ramakrishna, M., Richardson, A., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T., Span, P., Teague, J., Totoki, Y., Tutt, A., Valdés-Mas, R., van Buuren, M., van't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S., Siebert, R., Campo, E., Shibata, T., Pfister, S., Campbell, P. & Stratton, M. (2013), 'Signatures of mutational processes in human cancer', *Nature* **500**(7463), 415.

Anderson, A. (2005), 'A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion', *Math Med Biol* **22**(2), 163.

Anderson, A. & Chaplain, M. (1998), 'Continuous and discrete mathematical models of tumor-induced angiogenesis', *Bull Math Biol* **60**(5), 857–899.

Baker, A.-M., Cereser, B., Melton, S., Fletcher, A., Rodriguez-Justo, M., Tadrous, P., Humphries, A., Elia, G., McDonald, S. A., Wright, N., Simons, B., Jansen, M. & Graham, T. (2014), 'Quantification of crypt and stem cell evolution in the normal and neoplastic human colon', *Cell Rep* **8**(4), 940–947.

Bao, S., Wu, Q., McLendon, R., Hao, Y., Shi, Q., Hjelmeland, A., Dewhirst, M., Bigner, D. & Rich, J. (2006), 'Glioma stem cells promote radioresistance by preferential activation of the dna damage response', *Nature* **444**(7120), 756–760.

Bignell, G., Greenman, C., Davies, H., Butler, A., Edkins, S., Andrews, J., Buck, G., Chen, L., Beare, D., Latimer, C., Widaa, S., Hinton, J., Fahey, C., Fu, B., Swamy, S., Dalgliesh, G., Teh, B., Deloukas, P., Yang, F., Campbell, P., Futreal, P. & Stratton, M. (2010), 'Signatures of mutation and selection in the cancer genome', *Nature* **463**(7283), 893–898.

Blum, M. & François, O. (2005), 'On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited', *Math Biosci* **195**(2), 141–153.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A. et al. (2012), 'Absolute quantification of somatic dna alterations in human cancer', *Nature biotechnology* **30**(5), 413.

Chen, J., Li, Y., Yu, T., McKay, R., Burns, D., Kernie, S. & Parada, L. (2012), 'A restricted cell population propagates glioblastoma growth after chemotherapy', *Nature* **488**(7412), 522–6.

Conley, S., Gheordunescu, E., Kakarala, P., Newman, B., Korkaya, H., Heath, A., Clouthier, S. & Wicha, M. (2012), 'Antiangiogenic agents increase breast cancer stem cells via the generation of tumor hypoxia', *Proc Natl Acad Sci USA* **109**(8), 2784–2789.

Curtius, K., Wong, C., Hazelton, W., Kaz, A., Chak, A., Willis, J., Grady, W. & Luebeck, E. (2016), 'A molecular clock infers heterogeneous tissue age among patients with Barrett's esophagus', *PLoS Comput Biol* **12**(5), e1004919.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L. & Morris, Q. (2015), 'Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors', *Genome biology* **16**(1), 35.

Dhawan, A., Graham, T. & Fletcher, A. (2016), 'A computational modeling approach for deriving biomarkers to predict cancer risk in premalignant disease', *Cancer Prev Res* **9**(4), 283–295.

Dhawan, A., Kohandel, M., Hill, R. & Sivaloganathan, S. (2014), 'Tumour control probability in cancer stem cells hypothesis', *PLOS ONE* **9**(5), e96093.

Dhawan, A., Tonekaboni, S., Taube, J., Hu, S., Sphyris, N., Mani, S. & Kohandel, M. (2016), 'Mathematical modelling of phenotypic plasticity and conversion to a stem-cell state under hypoxia', *Sci Rep* **6**.

Diehn, M., Cho, R., Lobo, N., Kalisky, T., Dorie, M., Kulp, A., Qian, D., Lam, J., Ailles, L., Wong, M., Benzion, J., Kaplan, M., Wapnir, I., Dirbas, F., Somlo, G., Garbergolio, C., Paz, B., Shen, J., Lau, S., SR, Q., Brown, J., Weissman, I. & Clarke, M. (2009), 'Association of reactive oxygen species levels and radioresistance in cancer stem cells', *Nature* **458**(7239), 780–783.

Dingli, D. & Michor, F. (2006), 'Successful therapy must eradicate cancer stem cells', *Stem Cells* **24**(12), 2603–2610.

Doetsch, F., Petreanu, L., Caille, I., Garcia-Verdugo, J. & Alvarez-Buylla, A. (2002), 'EGF converts transit-amplifying neurogenic precursors in the adult brain into multipotent stem cells', *Neuron* **36**(6), 1021–1034.

Driessens, G., Beck, B., Caauwe, A., Simons, B. & Blanpain, C. (2012), 'Defining the mode of tumour growth by clonal analysis', *Nature* **488**(7412), 527–530.

Enderling, H., Anderson, A., Chaplain, M., Beheshti, A., Hlatky, L. & Hahnfeldt, P.

(2009), 'Paradoxical dependencies of tumor dormancy and progression on basic cell kinetics', *Cancer Res* **69**(22), 8814–8821.

Faltas, B., Prandi, D., Tagawa, S., Molina, A., Nanus, D., Sternberg, C., Rosenberg, J., Mosquera, J., Robinson, B., Elemento, O., Sboner, A., Beltran, H., Demichelis, F. & Rubin, M. (2016), 'Clonal evolution of chemotherapy-resistant urothelial carcinoma', *Nat Genet* **48**, 1490–1499.

Felsenstein, J. (1981), 'Evolutionary trees from dna sequences: a maximum likelihood approach', *J Mol Evol* **17**(6), 368–376.

Fialkow, P., Gartler, S. & Yoshida, A. (1967), 'Clonal origin of chronic myelocytic leukemia in man', *Proc Natl Acad Sci USA* **58**(4), 1468–71.

Flavahan, W., Wu, Q., Hitomi, M., Rahim, N., Kim, Y., Sloan, A. E., Weil, R., Nakano, I., Sarkaria, J., Stringer, B., Day, B. W., Li, M., Lathia, J., Rich, J. & AB, H. (2013), 'Brain tumor initiating cells adapt to restricted nutrition through preferential glucose uptake', *Nat Neurosci* **16**(10), 1373–1382.

Fletcher, W. & Yang, Z. (2009), 'INDELible: a flexible simulator of biological sequence evolution', *Mol Biol Evol* **26**(8), 1879–1888.

Gerlee, P. & Anderson, A. (2008), 'A hybrid cellular automaton model of clonal evolution in cancer: The emergence of the glycolytic phenotype', *J Theor Biol* **250**(4), 705–722.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A., Salm, M., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C., Martinex, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P., Stamp, G., Pickering, L., Gore, M., Nicol, D., Hazell, S., Futreal, P., Stewart, A. & Swanton, C. (2014), 'Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing', *Nat Genet* **46**(3), 225–233.

Gerlinger, M., Rowan, A., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C., Nohadani, M., Eklund, A., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. & Swanton, C. (2012), 'Intratumor heterogeneity and branched evolution revealed by multiregion sequencing', *N Engl J Med* **366**(10), 883–92.

Gilbertson, R. & Graham, T. (2012), 'Cancer: Resolving the stem-cell debate', *Nature* **488**(7412), 462–463.

Grenfell, B., Pybus, O., Gog, J., Wood, J., Daly, J., Mumford, J. & Holmes, E. (2004), 'Unifying the epidemiological and evolutionary dynamics of pathogens', *Science* **303**(5656), 327–332.

Heddleston, J., Li, Z., McLendon, R., Hjelmeland, A. & Rich, J. (2009), 'The hypoxic microenvironment maintains glioblastoma stem cells and promotes reprogramming towards a cancer stem cell phenotype', *Cell Cycle* **8**(20), 3274–84.

Hjelmeland, A., Wu, Q., Heddleston, J., Choudhary, G., MacSwords, J., Lathia, J., McLendon, R., Lindner, D., Sloan, A. & Rich, J. (2011), 'Acidic stress promotes a glioma stem cell phenotype', *Cell Death Differ* **18**(5), 829–840.

Hudson, R. R. (2002), 'Generating samples under a wright–fisher neutral model of genetic variation', *Bioinformatics* **18**(2), 337–338.

Kimura, M. (1969), 'The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations', *Genetics* **61**(4), 893.

Kirkpatrick, M. & Slatkin, M. (1993), 'Searching for evolutionary patterns in the shape of a phylogenetic tree', *Evolution* **47**(4), 1171–1181.

Lathia, J., Hitomi, M., Gallagher, J., Gadani, S., Adkins, J., Vasanji, A., Liu, L., Eyler, C., Heddleston, J., Wu, Q., Minhas, S., Soeda, A., Hoeppner, D., Ravin, R., McKay, R.,

McLendon, R., Corbeil, D., Chenn, A., Hjelmeland, A., Park, D. & Rich, J. (2011), 'Distribution of CD133 reveals glioma stem cells self-renew through symmetric and asymmetric cell divisions', *Cell Death Dis* **2**, e200.

Leventhal, G., Kouyos, R., Stadler, T., Von Wyl, V., Yerly, S., Böni, J., Cellerai, C., Klimkait, T., Günthard, H. & Bonhoeffer, S. (2012), 'Inferring epidemic contact structure from phylogenetic trees', *PLoS Comput Biol* **8**(3), e1002413.

Li, Z., Bao, S., Wu, Q., Wang, H., Eyler, C., Sathornsumetee, S., Shi, Q., Cao, Y., Lathia, J., McLendon, R., Hjelmeland, B. & Rich, J. (2009), 'Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells', *Cancer Cell* **15**(6), 501–13.

Liu, S., Ginestier, C., Ou, S. J., Clouthier, S., Patel, S., Monville, F., Korkaya, H., Heath, A., Dutcher, J., Kleer, C., Jung, Y., Dontu, G., Taichman, R. & Wicha, M. (2011), 'Breast cancer stem cells are regulated by mesenchymal stem cells through cytokine networks', *Cancer Res* **71**(2), 614–24.

Magee, J., Piskounova, E. & Morrison, S. (2012), 'Cancer stem cells: impact, heterogeneity, and uncertainty', *Cancer Cell* **21**(3), 283–296.

McFarland, C., Korolev, K., Kryukov, G., Sunyaev, S. & Mirny, L. (2013), 'Impact of deleterious passenger mutations on cancer progression', *Proc Natl Acad Sci USA* **110**(8), 2910–2915.

McFarland, C., Yaglom, J., Wojtkowiak, J., Scott, J., Morse, D., Sherman, M. & Mirny, L. (2017), 'The damaging effect of passenger mutations on cancer progression', *Cancer Res* **77**(18), 4763–4772.

Mir, A., Rosselló, F. & Rotger, L. (2013), 'A new balance index for phylogenetic trees', *Math Biosci* **241**(1), 125–136.

Mooers, A. & Heard, S. (1997), 'Inferring evolutionary process from phylogenetic tree shape', *Q Rev Biol* pp. 31–54.

Morton, C., Hlatky, L., Hahnfeldt, P. & Enderling, H. (2011), 'Non-stem cancer cell kinetics modulate solid tumor progression', *Theor Biol Med Mod* **8**, 48.

Murugaesu, N., Wilson, G., Birkbak, N., Watkins, T., McGranahan, N., Kumar, S., Abbassi-Ghadi, N., Salm, M., Mitter, R., Horswell, S., Rowan, A., Phillimore, B., Biggs, J., Begum, S., Matthews, N., Hochhauser, D., Hanna, G. & Swanton, C. (2015), 'Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy', *Cancer Discov* **5**(8), 821–831.

Naxerova, K. & Jain, R. (2015), 'Using tumour phylogenetics to identify the roots of metastasis in humans', *Nat Rev Clin Oncol* **12**(5), 258–272.

O'Connor, M., Xiang, D., Shigdar, S., Macdonald, J., Li, Y., Wang, T., Pu, C., Wang, Z., Qiao, L. & Duan, W. (2014), 'Cancer stem cells: a contentious hypothesis now moving forward', *Cancer Lett* **344**(2), 180–187.

Pang, R., Law, W., Chu, A., Poon, J., Lam, C., Chow, A., Ng, L., Cheung, L., Lan, X., Lan, H., Tan, V., Yau, T., Poon, R. & Wong, B. (2010), 'A subpopulation of CD26+ cancer stem cells with metastatic capacity in human colorectal cancer', *Cell Stem Cell* **6**(6), 603–15.

Poleszczuk, J., Hahnfeldt, P. & Enderling, H. (2014), 'Biphasic modulation of cancer stem cell-driven solid tumour dynamics in response to reactivated replicative senescence', *Cell Prolif* **47**(3), 267–276.

Poleszczuk, J., Hahnfeldt, P. & Enderling, H. (2015), 'Evolution and phenotypic selection of cancer stem cells', *PLoS Comput Biol* **11**(3), e1004025.

Ritsma, L., Ellenbroek, S., Zomer, A., Snippert, H., de Sauvage, F., Simons, B., Clevers, H. & van Rheenen, J. (2014), 'Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging', *Nature* **507**(7492), 362–365.

Rodriguez-Brenes, I., Komarova, N. & Wodarz, D. (2011), 'Evolutionary dynamics of feedback escape and the development of stem-cell-driven cancers', *Proc Natl Acad Sci USA* **108**(47), 18983–18988.

Roeder, I., Horn, M., Glauche, I., Hochhaus, A., Mueller, M. C. & Loeffler, M. (2006), 'Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications', *Nature medicine* **12**(10), 1181.

Rosen, D. (1978), 'Vicariant patterns and historical explanation in biogeography', *Syst Biol* **27**(2), 159–188.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A. & Shah, S. P. (2014), 'Pyclone: statistical inference of clonal population structure in cancer', *Nature methods* **11**(4), 396.

Sackin, M. (1972), '"Good" and "bad" phenograms', *Syst Biol* **21**(2), 225–226.

Scott, J., Dhawan, A., Hjelmeland, A., Lathia, J., Chumakova, A., Hitomi, M., Fletcher, A., Maini, P. & Anderson, A. (2019), 'Recasting the cancer stem cell hypothesis: unification using a continuum model of microenvironmental forces', *Curr Stem Cell Rep* .

Scott, J., Fletcher, A., Anderson, A. & Maini, P. (2016), 'Spatial metrics of tumour vascular organisation predict radiation efficacy in a computational model', *PLoS Comput Biol* **12**(1), e1004712.

Shao, K. & Sokal, R. (1990), 'Tree balance', *Syst Biol* **39**(3), 266–276.

Somarelli, J., Ware, K., Kostadinov, R., Robinson, J., Amri, H., Abu-Asab, M., Fourie, N., Diogo, R., Swofford, D. & Townsend, J. (2016), 'Phylooncology: Understanding cancer through phylogenetic analysis', *Biochim Biophys Acta* .

Sottoriva, A., Spiteri, I., Piccirillo, S., Touloumis, A., Collins, V., Marioni, J., Curtis, C., Watts, C. & Tavaré, S. (2013), 'Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics', *Proc Natl Acad Sci USA* **110**(10), 4009–4014.

Sottoriva, A., Verhoeff, J. J., Borovski, T., McWeeney, S., Naumov, L., Medema, J., Sloot, P. & Vermeulen, L. (2010), 'Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity', *Cancer Res* **70**(1), 46–56.

Sottoriva, Aand Vermeulen, L. & Tavaré, S. (2011), 'Modeling evolutionary dynamics of epigenetic mutations in hierarchically organized tumors', *PLoS Comput Biol* **7**(5), e1001132.

Sprouffske, K., AC, A., Radich, J., Carroll, M., Nedelcu, A. & Maley, C. (2013), 'An evolutionary explanation for the presence of cancer nonstem cells in neoplasms', *Evol Appl* **6**(1), 92–101.

Tamura, K., Aoyagi, M., Wakimoto, H., Ando, N., Nariai, T., Yamamoto, M. & Ohno, K. (2010), 'Accumulation of CD133-positive glioma cells after high-dose irradiation by Gamma Knife surgery plus external beam radiation', *J Neurosurg* **113**(2), 310–318.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Horswell, S., Chambers, T., OBrien, T., Lopez, J. I., Watkins, T. B., Nicol, D. et al. (2018), 'Deterministic evolutionary trajectories influence primary tumor growth: Tracerx renal', *Cell* **173**(3), 595–610.

Vermeulen, L., DeSousa, F., vander Heijden, M., Cameron, K., de Jong, J., Borovski, T., Tuynman, J., Todaro, M., Merz, C., Rodermond, H., Sprick, M., Kemper, K., Richel, D., Stassi, G. & Medema, J. (2010), 'Wnt activity defines colon cancer stem cells and is regulated by the microenvironment', *Nat Cell Biol* **12**(5), 468–76.

Vlashi, E., Lagadec, C., Vergnes, L., Matsutani, T., Masui, K., Poulou, M., Popescu, R., Della Donna, L., Evers, P., Dekmezian, C., Reue, K., Christofk, H., Mischel, P. & Pajonk, F. (2011), 'Metabolic state of glioma stem cells and nontumorigenic cells', *Proc Natl Acad Sci USA* **108**(38), 16062–7.

Werner, B., Dingli, D., Lenaerts, T., Pacheco, J. & Traulsen, A. (2011), 'Dynamics of mutant cells in hierarchical organized tissues', *PLoS Comput Biol* **7**(12), e1002290.

Werner, B., Scott, J., Sottoriva, A., Anderson, A., Traulsen, A. & Altrock, P. (2016), 'The cancer stem cell fraction in hierarchically organized tumors can be estimated using mathematical modeling and patient-specific treatment trajectories', *Cancer Res* **76**(7), 1705–1713.

White, J., Rassweiler, A., Samhouri, J., Stier, A. & White, C. (2014), 'Ecologists should not use statistical significance tests to interpret simulation model results', *Oikos* **123**(4), 385–388.

Williams, M., Werner, B., Barnes, C., Graham, T. & Sottoriva, A. (2016), 'Identification of neutral tumor evolution across cancer types', *Nat Genet* **48**, 238–244.

Yule, G. (1925), 'A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis, FRS', *Phil Trans R Soc B* **213**, 21–87.

Zhao, Z., Zhao, B., Bai, Y., Iamarino, A., Gaffney, S., Schlessinger, J., Lifton, R., Rimm, D. & Townsend, J. (2016), 'Early and multiple origins of metastatic lineages within primary tumors', *Proc Natl Acad Sci USA* **113**(8), 2140–2145.

Fig 1. Spatial stochastic model schematic with neutral mutation schema. (A) The proliferative hierarchy. Each TIC can divide symmetrically with probability α to make two identical TIC progeny, or asymmetrically with probability $(1-\alpha)$ to make one TIC and one TAC. TACs divide symmetrically until they reach a specific divisional age ($\beta = 4$ for this work), after which they die upon division attempt. (B) At each division event (branching) after the first (carcinogenesis, labelled with a 1), a random number of mutations drawn from a Poisson distribution with expectation λ is conferred on each daughter (subsequent starred events). Each mutation event is given a unique flag, which is inherited by its offspring unless they too mutate. Each unique mutation can then be considered as a novel mutant allele (red) appearing in the population. (C) Flowchart outlining cellular automaton rules governing TIC and TAC growth, including spatial inhibition of growth and TAC age.

Fig 2. Temporal evolution of the spatial model reveals observable morphologic differences between TIC-driven and non-TIC-driven tumours, as observed by others. We plot representative results of simulations of two tumours, each simulated on a square lattice of size 400×400. Top: a tumour simulated with $\alpha = 0.2$ and $\beta = 4$. We notice, as have Enderling et al. (2009) and Sottoriva et al. (2010), a 'patchy' clonal architecture, and non-uniform edge. Bottom: a tumour simulated with $\alpha = 1.0$, i.e. no proliferative hierarchy. We note smooth edges, radial patterns of clonal architecture and relatively faster population growth, reaching $\approx 70,000$ cells in less than 200 time steps. To reach a similar size, the tumour with symmetric division probability of 0.2 took $35,000$ time steps. Colour bars denote number of mutations present in a given clone, note that the top scale is about 1/3 of bottom scale.

Fig 3. Three example simulations with increasing symmetric division probability, α (0.2, 0.6 and 1.0 from top to bottom) and their associated phylogenetic trees. Each example plot is the result of a single stochastic simulation of our spatial CA model. Each simulation is initiated with a single TIC and complete when

the domain is full, in this case $250,000$ cells. Parameter values are $\beta = 4$ and $\lambda = 0.01$. Visualized trees (right) have been pruned of all leaves for ease of visualisation, which does not qualitatively affect measure rank (see Fig 8).

Fig 4. Example phylogenetic trees and their measures. From left to right the trees contain 4, 3 and 2 internal nodes (dots) respectively, but the same number (6) of terminal nodes.

Fig 5. A summary of four tree indices measured over a range of symmetric division probability. We plot the distribution of each of four measures of tree balance for the final resultant trees from 50 simulations against symmetric division probability. All simulations were run with $\beta = 4$ and $\lambda = 0.01$ until a tumour size of $250,000$ cells was reached. In each plot we display a box-whisker plot as well as the individual results as points. NS = non-significant by the Wilcoxon rank sum test.

Fig 6. Comparing phylogenetic tree measures across symmetric division probability through tumour growth. We plot the average and standard deviation (error bars) of four phylogenetic tree measures for each of the 50 simulations for a range of symmetric division probabilities over the course of tumour growth. Rank is maintained across symmetric division probabilities for each of the 3 tree measures with which we could discriminate between symmetric division probabilities. As before, $\bar{N}$ is not predictive and changes rank throughout tumour growth. All tumours are grown to eventual confluence at $250,000$ cells. In all simulations $\beta = 4$ and $\lambda = 0.01$.

Fig 7. Comparing phylogenetic tree measures across symmetric division probability and mutation probability. We plot the average of each of four phylogenetic tree measures at the end of each of 10 simulations for a range of symmetric division probabilities and mutation probabilities. We vary mutational probability over two orders of magnitude $(0.1 - 0.001)$, and simulate all tested symmetric division probabilities. Rank is maintained across symmetric division probabilities for each of the three of the four measures with which we could discriminate between symmetric division probabilities with

changing mutation probability, allowing for differentiation between parameters. As before, the $\bar{N}$ statistic is not predictive. As expected, for the non-normalized indices, Sackin and B1, the measures change monotonically with both symmetric division and mutation probability. For the PDA normalized Sackin index, however, there is a global minimum for $\lambda = 0.01$ and $\alpha = 1$.

Supplementary Fig 8. Raw and pruned trees give rise to qualitatively similar summary measures with rank preserved. For each tree-based measure considered in the main text, we plot the measure based on the full (upper) and pruned (lower) tree. For each pair, we plot the results from 10 simulations for each of the tested symmetric division probabilities. From left to right, we plot the B1 statistic, $\bar{N}$, the Sackin index, the PDA normalised Sackin index and finally the Yule normalised Sackin index.

Supplementary Fig 9. Effect size of symmetric division for four tree-based measures. We plot the effect size for the data shown in Fig 5.