A Computational Modelling Approach for Deriving Biomarkers to Predict Cancer Risk in Premalignant Disease

Andrew Dhawan[1], Trevor A. Graham[2,*], Alexander G. Fletcher[3,*],

1 School of Medicine, Queen's University, Kingston, Ontario, Canada

2 Barts Cancer Institute, Queen Mary University of London, London, UK

3 Mathematical Institute, University of Oxford, Oxford, UK

*** t.graham@qmul.ac.uk (TG); alexander.fletcher@maths.ox.ac.uk (AF)**

Abstract

The lack of effective biomarkers for predicting cancer risk in premalignant disease is a major clinical problem. There is a near-limitless list of candidate biomarkers and it remains unclear how best to sample the tissue in space and time. Practical constraints mean that only a few of these candidate biomarker strategies can be evaluated empirically and there is no framework to determine which of the plethora of possibilities is the most promising. Here we have sought to solve this problem by developing a theoretical platform for *in silico* biomarker development. We construct a simple computational model of carcinogenesis in premalignant disease and use the model to evaluate an extensive list of tissue sampling strategies and different molecular measures of these samples. Our model predicts that: (i) taking more biopsies improves prognostication, but with diminishing returns for each additional biopsy; (ii) longitudinally-collected biopsies provide only marginally more prognostic information than a single biopsy collected at the latest possible time-point; (iii) measurements of clonal diversity are more prognostic than measurements of the presence or absence of a particular abnormality and are particularly robust to confounding by tissue sampling; and (iv) the spatial pattern of clonal expansions is a particularly prognostic measure. This study demonstrates how the use of a mechanistic framework provided by computational modelling can diminish empirical constraints on biomarker development.

Introduction

Each year, tens of thousands of patients in the UK are diagnosed with a premalignant disease, a benign condition that predisposes to the future develop of cancer. Examples of common premalignant diseases include Barrett's Oesophagus [1], Ductal Carcinoma *in situ* (DCIS) of the breast [2], benign prostatic intraepithelial neoplasia (PIN) [3], and carcinoma *in situ* in the bladder [4]. The clinical management of patients with premalignant disease is a major challenge: in order to prevent cancer, patients are typically enrolled into longitudinal screening programmes that aim to detect (and then treat) patients who show early signs of progression to cancer. However, while having a premalignant disease increases the *average* risk of developing cancer compared to the unaffected population, the cancer risk for any individual is highly variable and generally quite low. For example, patients with Barrett's Oesophagus have an average 40-fold increased lifetime risk of developing adenocarcinoma, but the progression rate per patient per year is less than 0.5% [5] and so many of these patients will not progress to cancer in their lifetime. As a result, it is arguable that surveying an average (low-risk) patient is unnecessary as they are unlikely to ever progress to cancer. In addition, the surveillance process is typically unpleasant for the patient, and is very costly to health-care providers. In view of these facts together, premalignant disease is often described as both *over-diagnosed* and *over-treated* [6], and consequently there is a pressing clinical need to be able to accurately stratify cancer risk in these patients.

Prognostic biomarkers are central to current risk-stratification strategies. Here a biomarker is defined as an analysable property of the diseased tissue that correlates with the risk of progressing to cancer. In general, it remains unclear which of the plethora of potential biological features that could be assayed (morphological, gene expression, mutation, or other features) offers the most potential for prognostic value. Pathological grading and staging remain the most widespread biomarkers in current use; these biomarkers are descriptions of the morphlogical features of the disease. The current state-of-the-art biomarkers are molecular in nature, and typically quantify the aberrant expression of a panel of carefully-chosen genes. For example, the Oncotype DX assay analyses the activity of 21 genes to determine a score quantifying risk of recurrent breast cancer and response to chemotherapy [7]. Genetically based biomarkers include EGFR mutations in non-small cell lung cancer [8] and *TP53* abnormalities in Barrett's Oesophagus [9]. The limited predictive value of existing biomarkers has prevented their widespread clinical use [10], and for many diseases such as DCIS [11] and inflammatory bowel disease [12] no prognostic biomarkers have yet been identified.

All biomarkers require the diseased tissue to be sampled. Needle biopsies are the predominant sampling method, although other tissue collection methods such as endoscopic brushings or cell washings are sometimes used. However, typically the prognostic optimality of different sampling schemes, including whether samples should be collected longitudinally, has not been evaluated. Furthermore, given the fact that taking a biopsy is an invasive procedure, an empirical evaluation of different tissue sampling schemes is largely unfeasible.

Cancer development is fundamentally an evolutionary process: the acquisition of random somatic mutations can cause a cell to develop an evolutionary advantage over its neighbours, and so drive the clonal expansion of the mutant. Repeated rounds of mutation and clonal selection can lead to the development of a malignant tumour. When viewed from this evolutionary perspective, a biomarker may be thought of as a predictor of the *evolutionary trajectory* of the disease; a successful biomarker is one that sensitively and specifically detects which premaligant lesions are (rapidly) evolving towards cancer. However, existing biomarker development efforts do not explicitly consider the evolutionary process they seek to assay, instead relying on the identification of a small set of genes that are aberrantly expressed in high-risk cases [10]. The recent appreciation that carcinogenesis is a highly stochastic process [13], in which many different combinations of genetic alterations and gene expression changes contribute to the same malignant phenotypes, has led to doubts about the utility of such “candidate gene” approaches [14]. Alternative biomarker development strategies attempt to assay the underlying evolutionary *process* itself. Quantification of within-tumour diversity, as a proxy measure of the probability that the tumour has evolved a well-adapted “dangerous” clone, is one such measure that has shown efficacy in a variety of cancer types [15, 16, 17]. Whilst most studies have focused on the quantification of within-tumour genetic diversity, it is noteworthy that quantification of phenotypic heterogeneity also shows prognostic value [18, 19].

Mathematical models are tools that have the potential to diminish the inherent constraints of empirical biomarker development. Due to the relative ease with which a mathematical model of cancer evolution can be analysed, potentially exhaustive searches of candidate biomarkers can be performed *in silico*. This is the idea that we develop in this study.

Mathematical modelling has a rich history in cancer research, and is increasingly used as a tool to investigate and test hypothesized mechanisms underlying tumour evolution [20]. A common approach is to consider spatially homogeneous well-mixed populations [21], using multi-type Moran models of constant or exponentially growing size [22] or multi-type branching processes [23]. Other work has highlighted the impact of spatial dynamics on the evolutionary process [24]. More complex models have coupled a discrete representation of the movement and proliferation of individual cells to a continuum description of microenvironment factors such as oxygen concentration and extracellular matrix composition. Such models, in particular the pioneering work of Anderson and colleagues [25, 26], demonstrate the significant selective force imposed by microenvironmental conditions such as hypoxia. A recent discussion of the use of ecological and evolutionary approaches to study cancer is provided by Korolev et al. [27]. The majority of models of tumour evolution have focused on the rates of invasion and accumulation of mutations, and how these depend on factors such as modes of cell division and spatial heterogeneity in cell proliferation and death. Defining statistics that correlate with prognosis in these kinds of models is an unaddressed problem.

Here we use mathematical modelling as a novel platform for *in silico* biomarker development. We develop a simple mathematical model of tumour evolution, and use the model to evaluate the prognostic value of a range of different potential biomarker measures and different tissue sampling schemes.

Methods

Computational model of within-tumour evolution and biopsy sampling

To simulate the growth and dynamics of a pre-cancerous lesion, we consider a continuous-time spatial Moran process model of clonal evolution [28] on a two-dimensional square lattice, which may be thought of as a mathematical representation of an epithelial tissue. This description is similar to a model of field cancerization proposed by Foo et al. [29], although our model differs in several respects, which we describe below. We assume that in the transition from pre-malignant to malignant lesions, cells in a spatially well-structured population such as an epithelium are killed and/or extruded by an environmental stressor at a rate that is proportional to the inverse of their fitness, and replaced within the tissue via the division of a neighbouring cell. This assumption is represented in our chosen update rule. We suppose that it is this increased rate of cell turnover that leads to the accumulation of mutations, and eventually cancer. We refer to mutations as *advantageous*, *deleterious* or *neutral*, if they increase, decrease, or leave cell fitness unchanged.

The state of the system changes over time as a result of 'death-birth' events. At each point in time, each lattice site is defined by the presence of a cell with a specified 'genotype', given by the numbers of advantageous, neutral and deleterious mutations that it has accumulated. To implement the next death-birth event, we first choose a cell to die, at random, with a probability weighted by the inverse of each cell's fitness. We define the fitness of a cell with n_p advantageous, n_n neutral, and n_d deleterious mutations by

$$f = (1 + s_p)^{n_p}(1 - s_d)^{n_d}, \tag{1}$$

where the advantageous parameters s_p and s_d denote the relative fitness increase/decrease due to a advantageous/deleterious mutation. The chosen cell is removed from the lattice and one of the dead cell's neighbours is chosen uniformly at random to divide into the vacated lattice site. The time at which this death-birth event occurs is given by a waiting time, chosen according to an exponential distribution with mean equal to the sum of all cell inverse fitnesses present on the lattice, as stipulated by the Gillespie algorithm [30].

Immediately following division, each daughter cell can independently accrue a mutation, with probability μ. If a mutation is accrued, it is labelled as advantageous, deleterious or neutral with equal probability 1/3. We note that neutral and deleterious mutations are not typically included in spatial Moran models of tumour evolution, as such mutations are unlikely to persist. However, over shorter timescales, their presence may have an effect on the dynamics of the system and hence the predictive power of any biomarkers considered.

We define the *time to clinical detection* of cancer to be the earliest time at which the proportion of cells with at least N_m advantageous mutations exceeds a specified threshold δ. This reflects the time taken to reach a small, but clinically detectable, proportion of cancer cells that are capable of initiating and driving further tumour growth. In all simulations, we take $\delta = 0.05$. We evaluate the correlation of a measurement of some property of the state of the lesion sampled at some time T_b with the subsequent waiting time to cancer.

Measurements of the state of the lesion are performed by (i) taking a 'biopsy' from the lesion, and (ii) evaluating a putative 'biomarker assay' on the biopsy. Three different biopsy strategies are considered. First, we consider the whole lesion, in order to establish an upper bound on the prognostic power of each biomarker when using maximal information about the state of the system at a given time. Second, we sample a biopsy comprising a circular region of cells of radius N_b lattice sites, whose centre is chosen uniformly at random such that the entire biopsy lies on the lattice; this represents the clinical procedure of core needle sampling. Third, we sample N_s cells uniformly at random from the lattice; this represents washing or mechanical scraping of the lesion. In each case, we suppose that the biopsy constitutes a 'snapshot', and do not remove the sampled cells from the tissue. This simplifying assumption avoids the need to explicitly model the tissue response to wounding. The various biomarker assays evaluated are detailed below.

The definitions and values of all model parameters are summarized in Table 1. A MATLAB implementation of our model and simulated biopsy analysis is provided (see Software S1).

Classical biomarkers

Proportion of cells with at least two advantageous mutations. A commonly used class of biomarkers measure the proportion of cells in a biopsy staining positive for a given receptor. Examples include the estrogen receptor (ER), progesterone receptor (PR) and HER2/neu

Table 1. Parameter values used in the model.

Parameter	Description	Value(s)
δ	Detectable fraction of cancer cells in the tissue	0.05
N_m	Minimum number of advantageous mutations for cancer	$\{3, 5, 10, 15\}$
s_p	Fitness increase from a advantageous mutation	$\{0, 0.002, 0.02, 0.2\}$
s_d	Fitness decrease from a deleterious mutation	$\{0, 0.002, 0.02, 0.2\}$
μ	Probability of mutation per cell division	$\{0.01, 0.05, 0.1\}$
t_w	Time over which a cell stains positive for a recent mitosis	0.01
N	Number of cells in lattice	100×100
N_b	Radius of biopsy region	$\{5, 20, 40\}$
N_s	Number of cells taken in scraping	1000
T_b	Time at which sample is taken	$\{50, 80\}$

amplification staining commonly performed for malignancies of the breast [31, 32, 33]. Such assays are cost-effective and relatively simple to implement. In this work, we use the cutoff of a cell having acquired at least two advantageous mutations to be representative of a cellular change that is observable in this manner. The measure is calculated simply by the number of cells having at least two advantageous mutations ($n_p > 1$), divided by the total number of cells sampled.

Mitotic proportion. The proliferative fraction is often assayed with prgonostic intent. Ki-67 staining, which identifies those cells in active cell cycle, is often used to identify the proliferative fraction. To represent such a measure within the framework of our computational model, we fix a small time window t_w within which we consider a cell to be proliferating: thus a cell in the model stains positive for Ki-67 if it had proliferated in the previous t_w time units. The mitotic proportion at a given time t is defined as the number of cells undergoing mitosis in the time window $(t - t_w, t]$ divided by the number of cells in the lattice, N.

Measures of heterogeneity

Shannon index. The Shannon index H measures diversity among a population comprising different types [34]. For a population of K distinct types, each comprising a proportion p_k of the population, the Shannon index is defined as

$$H = -\sum_{k=1}^{K} p_k \log p_k. \tag{2}$$

To calculate H we define p_k such that each distinct triplet of advantageous, neutral and deleterious mutations is associated with a distinct clone within the model, and p_k represents the proportion of cells in this clone.

Simpson index. Another established measure of diversity is the Simpson index, S, defined as an average of each type in the population, weighted by its own proportion [35]. The Simpson index may be thought of as the probability that two randomly chosen members from the population are of the same type. To ensure that a higher value corresponds to greater diversity, we redefine S by instead taking the difference between this index and 1. Therefore for a population of K distinct types (defined as described for the Shannon index), each comprising proportion p_k of the population, we have

$$S = 1 - \sum_{k=1}^{K} p_k^2. \tag{3}$$

The Simpson index may be thought of as a measure of species dominance, in that as the index increases towards the maximum of 1, the evenness of the distribution of the population over the various types becomes increasingly skewed toward one type.

Moran's *I*. Moran's I is a measure of global spatial autocorrelation which computes a weighted statistical average of the deviation between data points in a set, weighted by their spatial distance [36]. Moran's I takes values in $[-1, 1]$. For a given set of values $\{X_1 \ldots X_N\}$, with mean $\overline{X}$, and spatial weight matrix $(w_{ij}) \in \mathbb{R}_+^{N\times N}$, Moran's I is defined as

$$I = \left(\frac{N}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}} \right) \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_{i=1}^{N} (X_i - \overline{X})^2} \right). \tag{4}$$

We take X_i to be the sum of the numbers of advantageous, neutral and deleterious mutations accumulated by the cell at lattice site i. Clinically, X_i may be thought of as a binary variable (0 or 1) indicating whether a given cell bears some dectable abnormality (for instance a particular number of advantageous mutations). The spatial weight matrix (w_{ij}) can be specified in several ways; here, we define

$$w_{ij} = \frac{1}{1 + d_{ij}}, \tag{5}$$

where d_{ij} is the Euclidean distance between the lattice sites indexed by $i, j \in \{1, \ldots, N\}$. With this functional form, neighbouring points that are closer together are weighted more heavily, thus contributing more to the measure.

Geary's C. Geary's C, like Moran's I, is a global measure of spatial autocorrelation. Geary's C takes values in $[0, 2]$, with higher values indicating less spatial autocorrelation, and lower values indicating a greater degree of spatial autocorrelation [37]. While Moran's I is a more global measurement and sensitive to extreme observations, Geary's C is more sensitive to differences in local neighbourhoods. For a given set of values $\{X_1 \dots X_N\}$, with mean $\overline{X}$, and a given spatial weight matrix $(w_{ij}) \in \mathbb{R}_+^{N\times N}$, Geary's C is defined as

$$C = \left(\frac{N-1}{2\sum_{i=1}^N \sum_{j=1}^N w_{ij}} \right) \left(\frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(X_i - X_j)^2}{\sum_{i=1}^N (X_i - \overline{X})^2} \right). \tag{6}$$

Here, our definitions of X_i and (w_{ij}) follow those given for Moran's I.

Index of positive proliferation (IPP). We next define a novel measure, termed the index of positive proliferation (IPP), that is a spatially weighted average of the location of mitotic cells and the number of advantageous mutations accrued by nearby cells. The biological motivation for this measure is to detect the recent clonal expansions of advantageous mutants, in order to quantify evidence of recent progression towards cancer. These observations imply that a situation in which proliferation is concentrated to regions of known increased numbers of advantageous mutations, is prognostically worse than a situation where proliferation is concentrated in regions of low numbers of advantageous mutations.

We define the IPP as follows. Consider a population of N cells, with the individual cells labelled as $X_1, \dots, X_N$. Suppose that a subset of these cells, $Y_1, \dots, Y_Q$, are proliferating at a given time. We define *cellular weights* $f_1, \dots, f_M \in \mathbb{R}^+$ as values such that a higher weight corresponds to a cellular state genetically closer to that of cancer. Thus, for a given spatial weight matrix $(w_{ij}) \in \mathbb{R}^{N\times N}$ we define the IPP as

$$\mathrm{IPP} = \frac{\sum_{i=1}^Q \sum_{j=1}^N w_{ij} f_j}{\sum_{i=1}^Q \sum_{j=1}^N w_{ij}}. \tag{7}$$

We define the weights w_{ij} as in equation (5), where d_{ij} is the Euclidean distance between cell X_i and proliferating cell Y_j, such that $X_i \neq Y_j$. In the case that $X_i = Y_j$, we take $w_{ij} = 0$.

For each cell i with m_i advantageous mutations, we define the cellular weight f_i as

$$f_i = \begin{cases} 0: & m_i < N_m - 2, \\ 1: & m_i \geq N_m - 2, \end{cases} \tag{8}$$

Since time steps correspond to ‘generations’ in our mathematical model, we store the locations of the most recent cell divisions within a given time window t_w, and regard these locations as locations that would stain positive for mitotic activity (e.g. via Ki-67 staining), to model the non-instantaneous process of detecting active cell division. Clinically, these cellular weights correspond to cells that are genotypically closer to the end state of cancer, and represent either cutoff points that may be detected by gene sequencing, or immunohistochemical changes.

Index of non-negative proliferation (INP). To model the case where it may not be feasible to observe an accumulation of advantageous mutations only, in the sense that mutations accumulated may be neutral as well, we define an additional measure termed the index of non-negative proliferation (INP). This measure is defined analogously to the IPP, but with the cellular weights chosen such that

$$f_i = \begin{cases} 0 : & m_i + n_i < \left\lfloor \frac{N_m}{2} \right\rfloor, \\ 1 : & m_i + n_i \geq \left\lfloor \frac{N_m}{2} \right\rfloor, \end{cases} \tag{9}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. Here, the sum of the number of advantageous and neutral mutations is considered to be the observable quantity, simulating a situation in which the observable information encapsulates and may skew the perception of the true genotypic state of the system.

Statistical methods

The predictive value (correlation with the waiting time to detectable cancer) of each sampling strategy and putative biomarker assay was evaluated using Kaplan-Meier curves and univariate Cox Proportional Hazards models as implemented in the *R* statistical computing language.

Results

We consider a spatial model of the evolution of malignancy in a precancerous lesion. In our model, cells occupy a two-dimensional lattice of size N. The model is simulated as a succession of discrete time steps, with waiting times determined by a Gillespie algorithm [30]. At each time step, a cell is chosen at random to die and is removed from the lattice with a probability that is inversely proportional to its *cellular fitness*, a positive real number that is initially equal to 1 for non-mutated cells and may be altered by mutation. When a cell dies, one of its neighbours is then chosen uniformly at random to divide, with one of the daughter cells occupying the free lattice site and each daughter cell independently acquiring a new mutation with probability μ. We refer to mutations as *advantageous*, *deleterious* or *neutral*, according to whether they increase, decrease, or leave fitness unchanged, with each type of mutation assumed to be equally likely.

Starting from a lattice occupied entirely by non-mutant cells, we consider the *outcome* of each simulation to be the time taken for the proportion of cells with at least N_m advantageous mutations to exceed a threshold δ. This waiting time is defined as the time to clinically detectable lesions. We choose a value of δ corresponding to a proportion of cancer cells that is sufficiently large to be clinically detectable, and to initiate subsequent rapid growth.

A representative snapshot of a model simulation is shown in Fig. 1A. The model exhibits successive clonal sweeps of mutations (Fig. 1C). To simulate clinical sampling, at a predetermined time T_b we take a virtual *biopsy* from the lesion (Fig. 1B), from which we compute various biomarkers and assess their prognostic value in determining the waiting time to cancer (see *Methods*).

Assessment of candidate biomarkers and tissue sampling schemes

Small needle biopsies. We computed the prognostic value of various candidate biomarker 'assays' performed on a single biopsy of radius $N_b = 20$ taken at time T_b post simulation initiation. Neither the proportion of cells with at least one advantageous mutation nor the proliferative fraction were significant predictors of prognosis (Table 2). In contrast, measures of clonal diversity (Shannon and Simpson index) were both highly significant predictors of prognosis ($p < 10^{-4}$ in both cases). Of the spatial autocorrelation measures, Moran's I ($p = 0.02$) but not Geary's C ($p = 0.29$) had prognostic value.

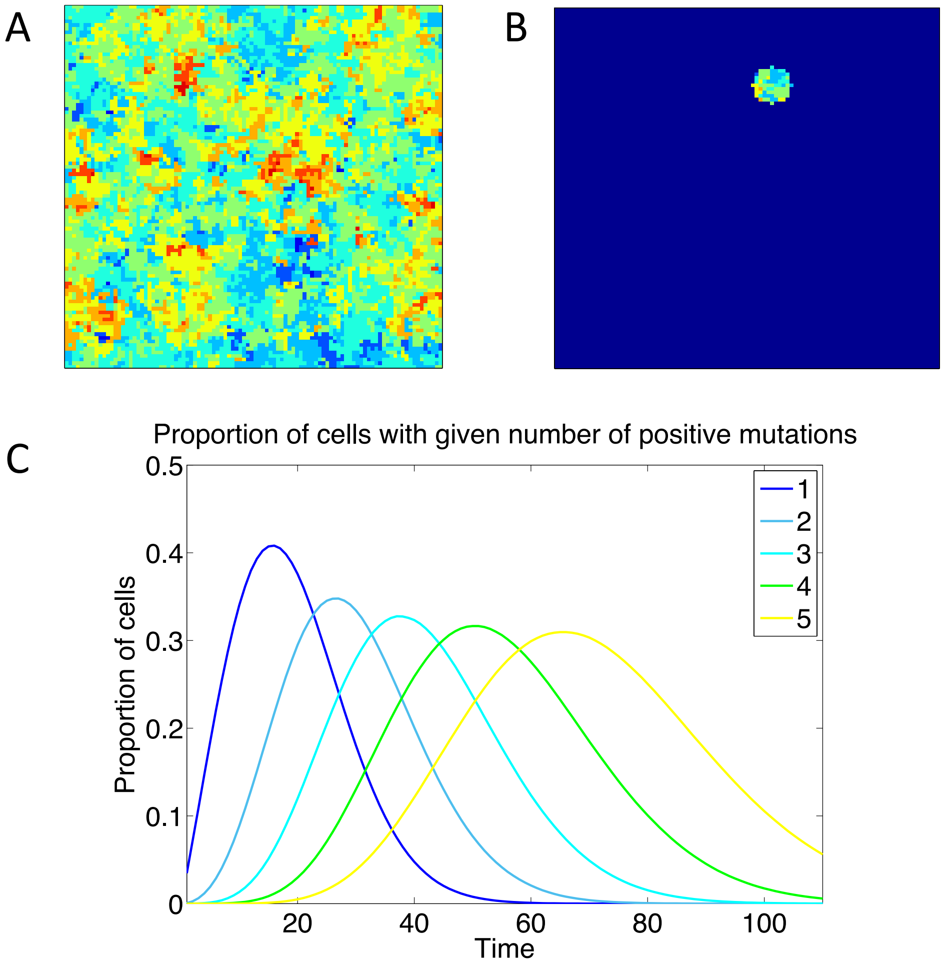


Figure 1. Depiction of the spatial simulation, a virtual biopsy, and the successive clonal sweeps. A: Heat map of the lattice at a given point in time, with different colours representing different numbers of positive mutations of the cells at those points. B: Depiction of the lattice subset involved in a virtual biopsy. C: Time evolution of the proportions of cells with different numbers of positive mutations, showing successive clonal sweeps. Results are shown from simulations with parameter values $N_m = 10$, $s_p = s_d = 0.2$.

Random sampling. Random sampling of cells from the lesion represents a tissue collection method such as an endoscopic brush or a cellular wash. We took a random sample of 10^3 cells, corresponding to 10% of the total lesion. As for small biopsy sampling, the proportion of mitotic cells within the sample was a poor prognostic marker ($p = 0.23$), but interestingly the proportion of cells with more than one advantageous mutation became a significant predictor ($p = 0.01$). This may be due to the fact that within a sparse sample of the lesion, the number of

Table 2. Summary of Cox proportional hazards models for various putative biomarker schemes, for different tissue sampling schemes. Hazard ratios (HR75) are computed at time $t = 75$ for the case $N_m = 10$, $s_p = s_d = 0.2$, and $\mu = 0.1$. Statistically significant values are in bold.

		Unit change	HR75	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.05	0	$(0, \infty)$	0.29
	Mitotic proportion	0.01	**0.31**	(0.12, 0.81)	0.02
	Shannon index	0.1	**2**	(1.8, 2,2)	$< 10^{-4}$
	Simpson index	0.01	**5.5**	(4.2, 7.2)	$< 10^{-4}$
	Moran's I	0.05	**3.2**	(2, 5.3)	$< 10^{-4}$
	Geary's C	0.01	0.98	(0.92, 1)	0.43
	IPP	0.01	**2**	(1.9,2.1)	$< 10^{-4}$
	INP	0.01	**1.3**	(1.2, 1.4)	$< 10^{-4}$
Biopsy	$n_p > 1$ proportion	0.05	0.95	(0.86, 1.1)	0.35
	Mitotic proportion	0.01	0.81	(0.58, 1.1)	0.22
	Shannon index	0.2	**1.3**	(1.1,1.4)	$< 10^{-4}$
	Simpson index	0.01	**2.3**	(1.5, 3.5)	$< 10^{-4}$
	Moran's I	0.1	**1.4**	(1.1, 1.9)	0.02
	Geary's C	0.1	0.95	(0.87,1)	0.29
	IPP	0.01	**1.1**	(1, 1.1)	$< 10^{-4}$
	INP	0.1	1	(0.95, 1.1)	0.57
Scraping	$n_p > 1$ proportion	0.05	$< 10^{-6}$	(0, 0.0001)	0.01
	Mitotic proportion	0.01	1.2	(0.88, 1.7)	0.23
	Shannon index	0.05	**1.3**	(1.3, 1.4)	$< 10^{-4}$
	Simpson index	0.01	**3.5**	(2.8, 4.4)	$< 10^{-4}$

mutant cells is a proxy for active on-going evolution: either via the large scale clonal expansion of a single clone, or multiple foci of independent clones. Increased clonal diversity remained a highly significant predictor of a short waiting time to cancer ($p < 10^{-4}$ for both the Shannon and Simpson indices; Fig. 2).

Whole lesion sampling. In the case of whole-lesion sampling, all information on the current state of the virtual tumour is available in the biomarker assay, and hence we expected to see maximum perdictive value of our putative biomarkers. In this case, the proportion of cells with at least one advanatageous mutation remained a poor prognosticator ($p = 0.29$), whereas the proportion of proliferative cells became a significant predictor ($p = 0.02$) (see Table 2).

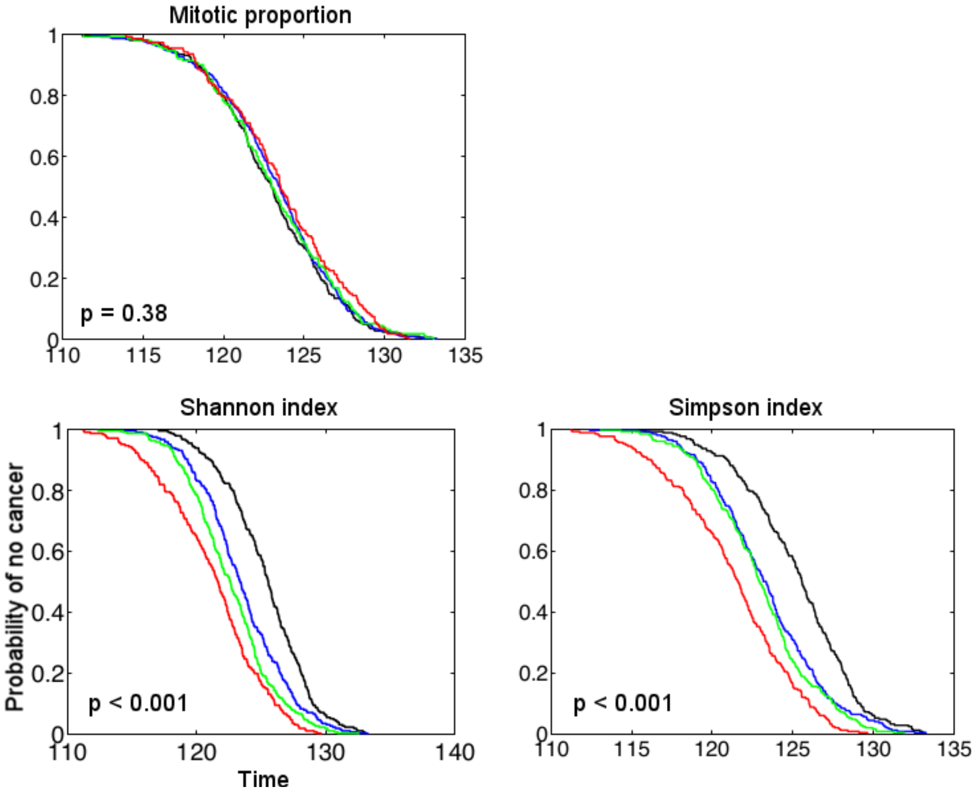


Figure 2. Prognostic value of random tissue sampling. A random sample of random sample of $N_s = 10^3$ (10% of the lesion) cells was sampled at time $T_b = 80$ and the prognostic value of the mitotic proportion (A), Shannon index (B) and Simpson index (C) on this sample was considered. Kaplan-Meier curves are plotted for each putative biomarker assessed, and in case, the values across the simulations were separated into upper (red), upper middle (green), lower middle (blue) and lower (black) quartiles. Only biomarkers that did not require spatial information could be computed for this tissue sampling method. *P*-values are for the generalized log-rank test.

The clonal diversity measures remained highly significant prognosticators ($p < 10^{-4}$ in both cases), underlining their robustness as prognostic measures. Higher clonal diversity was associated with faster progression to cancer (Fig. 3). The prognostic value of the spatial autocorrelation measure Moran's *I* was significantly improved when the whole grid was sampled ($p < 10^{-4}$), but Geary's *C* remained non-predictive.

Together these data highlight the high prognostic value of diversity measures, and their robustness to the details of tissue sampling method used.

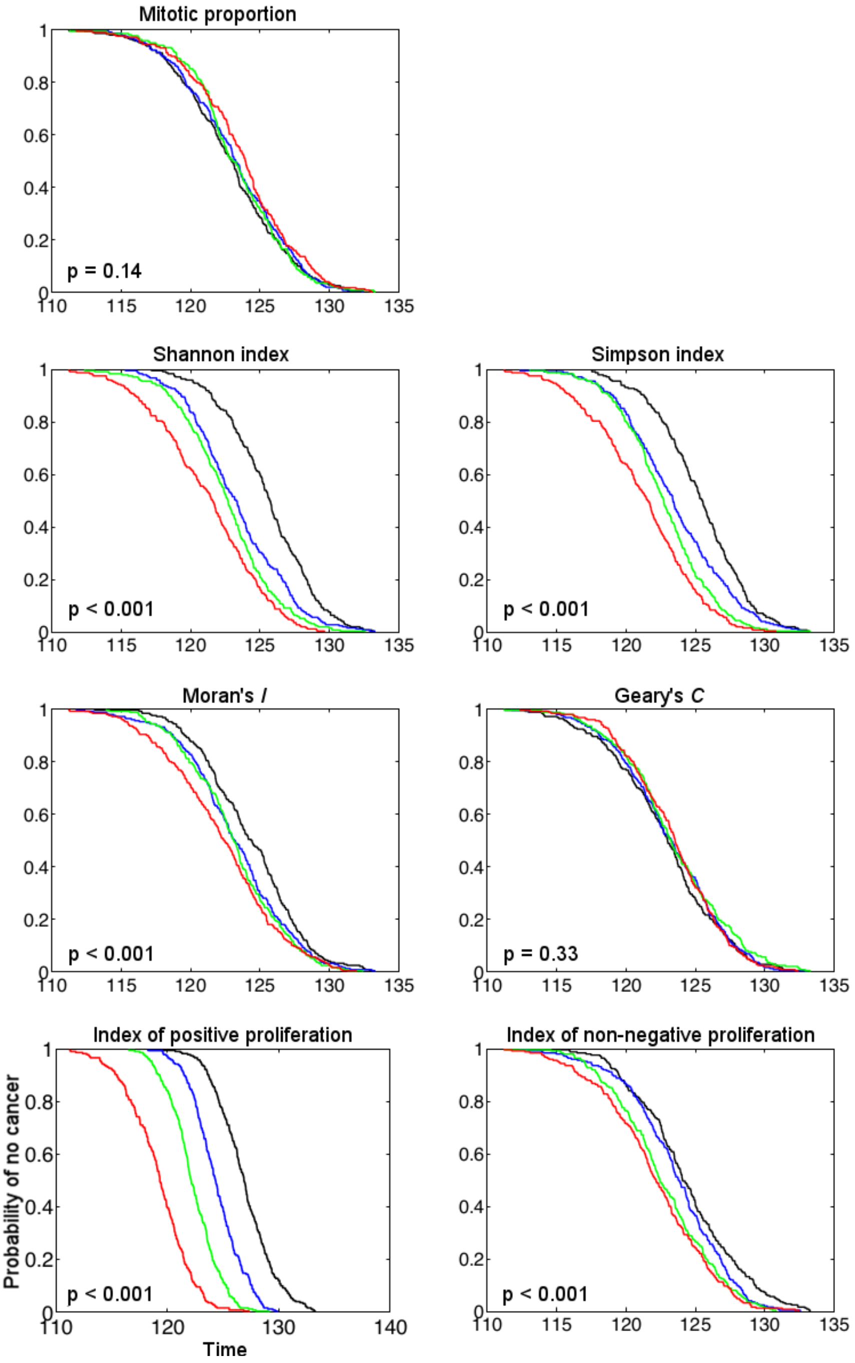


Figure 3. Sampling the whole lesion improves the prognostic value. The prognostic value of sampling the whole lattice at time $T_b = 80$ was assessed. Kaplan-Meier curves are plotted for each putative biomarker assessed for biomarker values across the simulations were separated into upper (red), upper middle (green), lower middle (blue) and lower (black) quartiles. *P*-values are for the generalized log-rank test.

Novel prognostic measures

We next sought to determine whether novel statistics calculated on the state of the lesion could provide additional prognostic value. We defined two new statistics, the *index of positive proliferation* (IPP) and the *index of non-negative proliferation* (INP), which describe the spatial autocorrelation between proliferating cells with advantageous mutations, or proliferating cells with non-deleterious mutations, respectively (see *Methods*). Since these statistics tie together measures of both the mutation burden and proliferative index, we consider them to be measures of the degree of ‘evolutionary activity’.

In both small biopsy samples and whole-lesion analysis, the IPP was a highly prognostic statistic (Table 2), with larger values of the statistic accurately predicting shorter times to cancer (Fig. 3). The INP was prognostic on whole-lesion analysis (Fig. 3), but not on targeted biopsies (Table 2). The difference in the prognostic value between IPP and INP is suggestive of the particular importance of assaying ‘distance’ travelled along the evolutionary trajectory towards cancer: the IPP is sensitive to this distance as it only measures advantageous mutations, whereas the INP is potentially confounded by non-adaptive mutations. The inherent issues associated with the identification of advantageous mutations consequently potentially limit the utility of these novel measures.

Early versus late biopsy

Effective screening for cancer risk requires predicting cancer risk long before the cancer develops. We next considered how the timing of a biospy affects its prognostic value by investigating how the correlation coefficient between each biomarker and the subsequent waiting time to cancer varies with the time at which the biopsy is taken.

As expected, we found that biopsies collected later in the lesion’s evolution (e.g. closer to the time of cancer development) generally had more predictive value than biopsies collected earlier, and this was true irrespective of the tissue sampling method used (Fig. 4A-C). Sampling early in the lesion’s evolution (e.g. near to the start of the simulations) had poor predictive value irrespective of the putative biomarker assay used, reflecting the fact that very few mutations had accumulated in the lesion at short times. Sampling at intermediate times showed dramatic improvements in the prognostic value of the diversity indices and IPP measure, whereas samples taken at a variety of long times had approximately equal prognostic value or showed slight declines relative to intermediate times. At intermediate and long times, the IPP was the best performing prognostic measure. The mitotic proportion and proportion of cells with at least one advantageous mutation were consistently poor predictors across the entire time course.

The effect of taking a small biopsy, as opposed to sampling the whole lesion, was to both significantly reduce the prognostic value of all putative biomarker measures, and introduce ‘noise’ into their prognostic values (Fig. 4B). Importantly, we observed that in spite of this noise, the correlation coefficients for the clonal heterogeneity and IPP measures were consistently high compared to the other measures, indicating their robustness as prognostic markers. Biopsy sampling significantly reduced the prognostic value of Moran’s *I* compared to whole-lesion sampling, indicating how this measure is particularly confounded by tissue sampling.

On random samples (analagous to endoscopic brushings or washings), the Shannon and Simpson indices showed good correlations with the waiting time to cancer. These diversity measures were more predictive for random samples than they were for circular biopsies, even though both methods sampled similar numbers of cells (10% of the lesion for brushings versus 12% for biopsies). This result may reflect the fact that a biopsy sample can potentially miss a ‘dangerous’ clone, whereas a random sampling method is likely to obtain cells from all sizeable clones within the lesion.

Together, these data indicate that larger samples usually provide more prognostic value than smaller samples, and that very ‘early’ tissue samples are unlikely to contain significant prognostic information. They also highlight again that the prognostic value of diversity measures is particularly robust to the details of tissue sampling.

Longitudinally collected biopsies

We next examined whether combining information from serial biopsies, taken at two different time points (t_1 and t_2; both strictly before cancer occurrence) provided more prognostic information that a biopsy from only a single time point. We took biopsies at two different time points and computed our putative biomarker scores for each time-point. We then calculated the difference in biomarker scores between the two time points and determined the correlation between this difference and the waiting time to cancer. To determine whether serial biopsies provided more prognostic information than a single time point, we determined whether or not the difference in biomarker values between time points was more or less correlated with the waiting time to cancer than the biomarker value at t_2 alone. These results are shown in Fig. 5, where the x and y axes indicate the time of the first and second biopsies, and the colours indicate the difference between the correlation coefficient for the *difference* of the individual biomarker values at each time points and the correlation coefficient for the biomarker value at the second time point alone.

Surprisingly, including information from an early biopsy provided little or no additional prognostic value over-and-above the information available in the later biopsy (Fig. 5). Further, the prognostic value of the diversity measures (Shannon and Simpson indices) was decreased by considering information from multiple time-points.

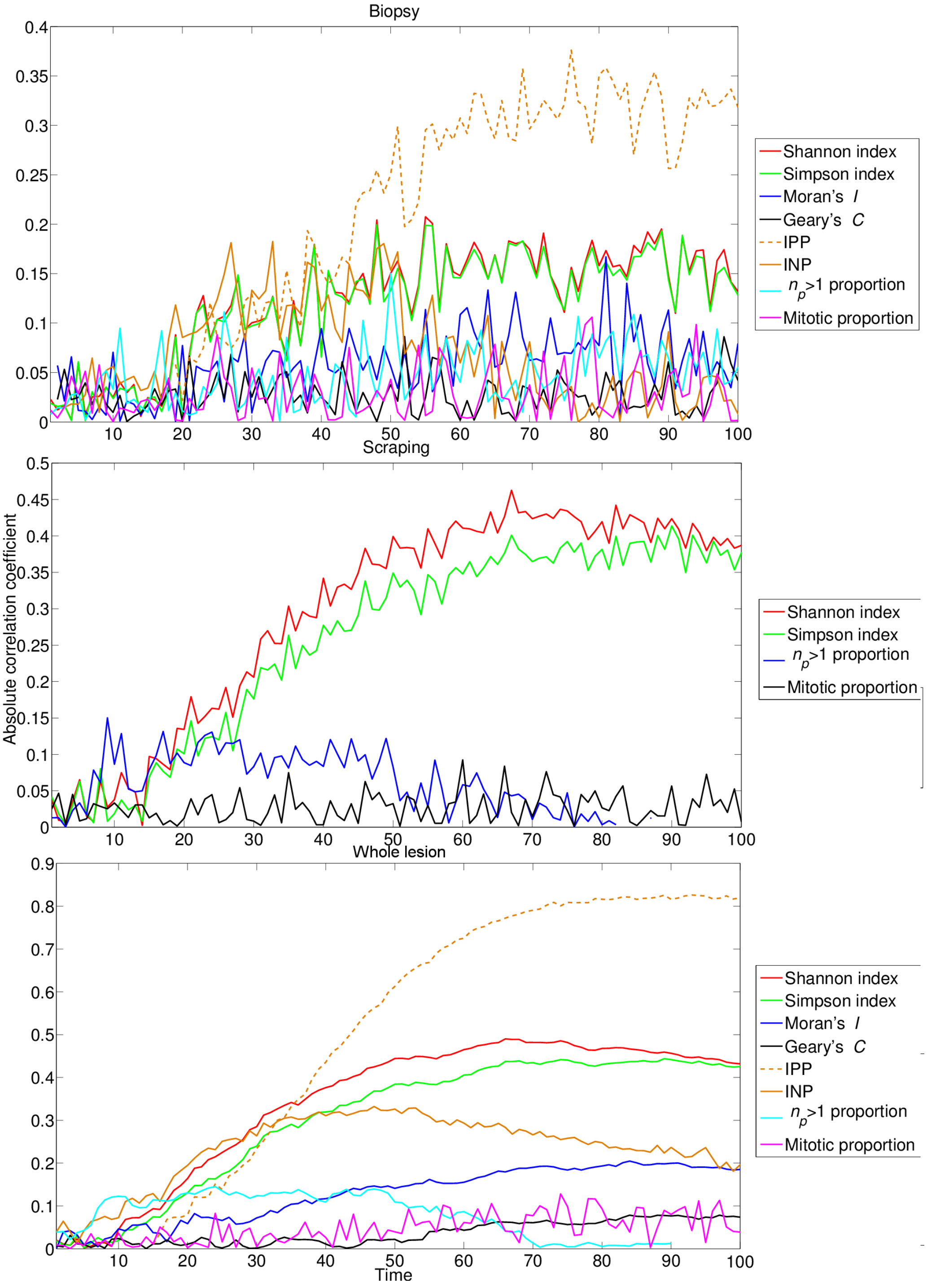


Figure 4. Prognostic value of early versus late biopsies. For a range of sampling times T_b, the virtual tissue was biopsied and the correlation between putative biomarker values and the waiting time to cancer was computed. Results from 1000 simulations for each sampling scheme, with $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$, $N_b = 20$ and $N_s = 10^3$.

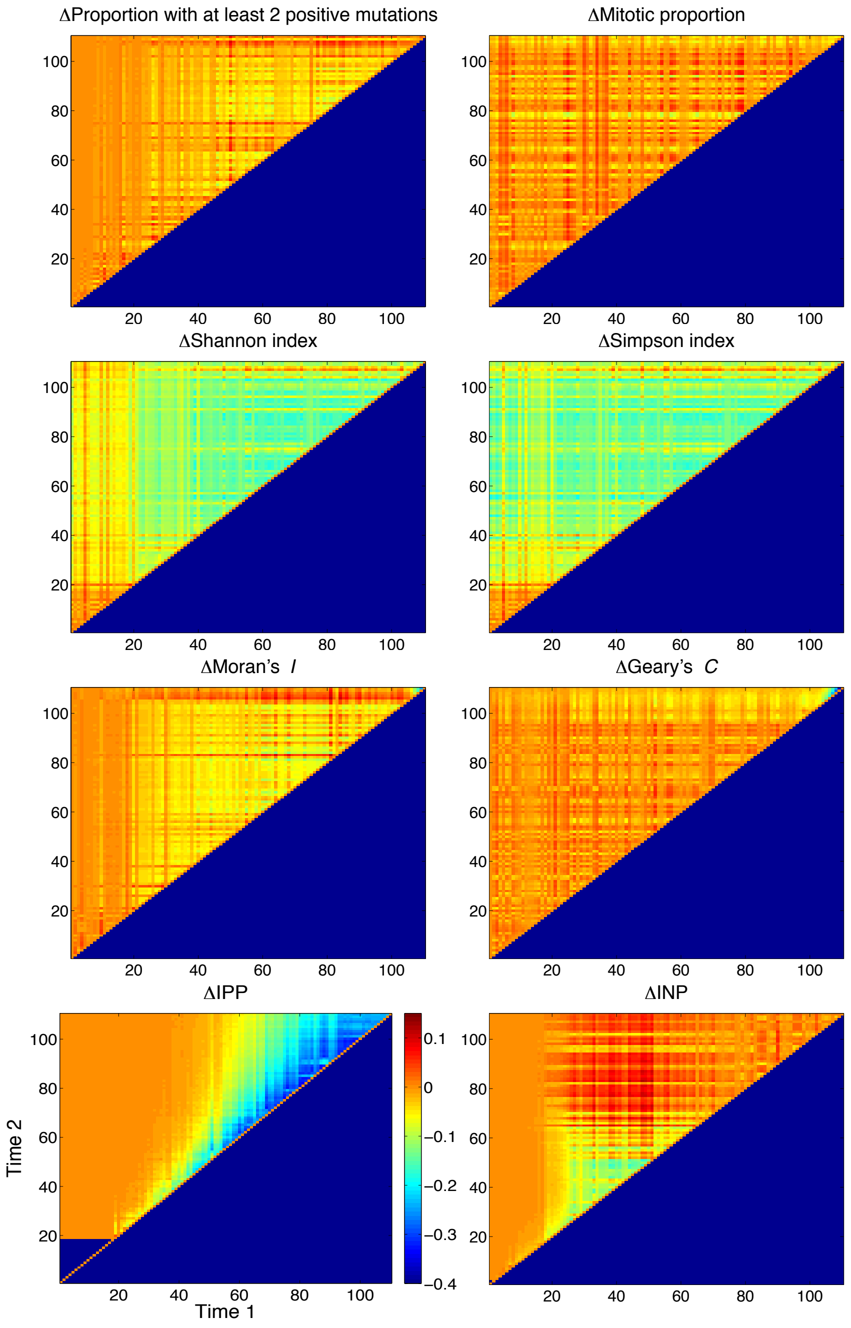


Figure 5. Serial biopsies provide only marginal additional prognostic information. Heat maps depicting the relative value of taking serial biopsies at different time points. Positive values (warm colours) indicate that prognostic value was improved by combining information from both time-points; negative values (cool colours) indicate that more information was available at the second time point along than for the combined time points. Results from 1000 simulations for for each pair of time points, with $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$ and $N_b = 20$.

Multiple biopsies at the same time point

A consequence of intra-tumour heterogeneity is that a single biopsy may fail to sample an important clone [38] and so cause an incorrect prognosis assignment. To address this issue, we studied how the prognostic value of our various putative biomarkers was improved by taking additional biopsies at the same time ($T_b = 50$). We emphasize that, for simplicity, in our virtual biopsy and biomarker assay the sampled tissue was perfectly replaced in order to avoid the complexities associated with modelling local wound healing and tissue recovery.

Taking more biopsies generally improved the prognostic value of all putative biomarker assays, but with diminishing returns for each additional biopsy (Fig. 6). With the exception of the proportion of cells with more than one advantageous mutation, the maximum prognostic value for each measure was achieved by taking the average biomarker score across all biopsies collected, whereas measures of the spread of values (either the variance or range of values) were generally poorer prognosticators. Interestingly, the maximum prognostic value for the proportion of cells with more than one advantageous mutation was achieved by taking the *minimum* value across the biopsies. This could be because this minimum value is particularly sensitive to biopsies that contain non-progressed cells. Together these data imply that taking more biopsies and averaging the biomarker signal across biopsies provides additional prognostic information.

Robustness of results to choice of model

To assess the robustness of results to our model assumptions, we investigated the impact of parameter values and update rules on the predictive value of each biomarker. We observed the same qualitative behaviour, such as diversity measures outperforming the proliferative fraction in predictive value, irrespective of the choice of parameter values or update rule used (see Figs S1–S3, Tables S1–S9 and Text S1 for details).

Discussion

In this work we have developed a simple computational model of cancer development within premalignant disease and used the model to evaluate the prognostic value of a range of different putative biomarker measurements and tissue sampling schemes. Our results show that simply counting the proportion of cells bearing multiple advantageous mutations (proportion of cells with ‘driver’ mutations) or the proportion of proliferating cells were universally poor prognostic biomarkers, whereas measures of clonal diversity were highly prognostic and robust to the choice of tissue sampling scheme.

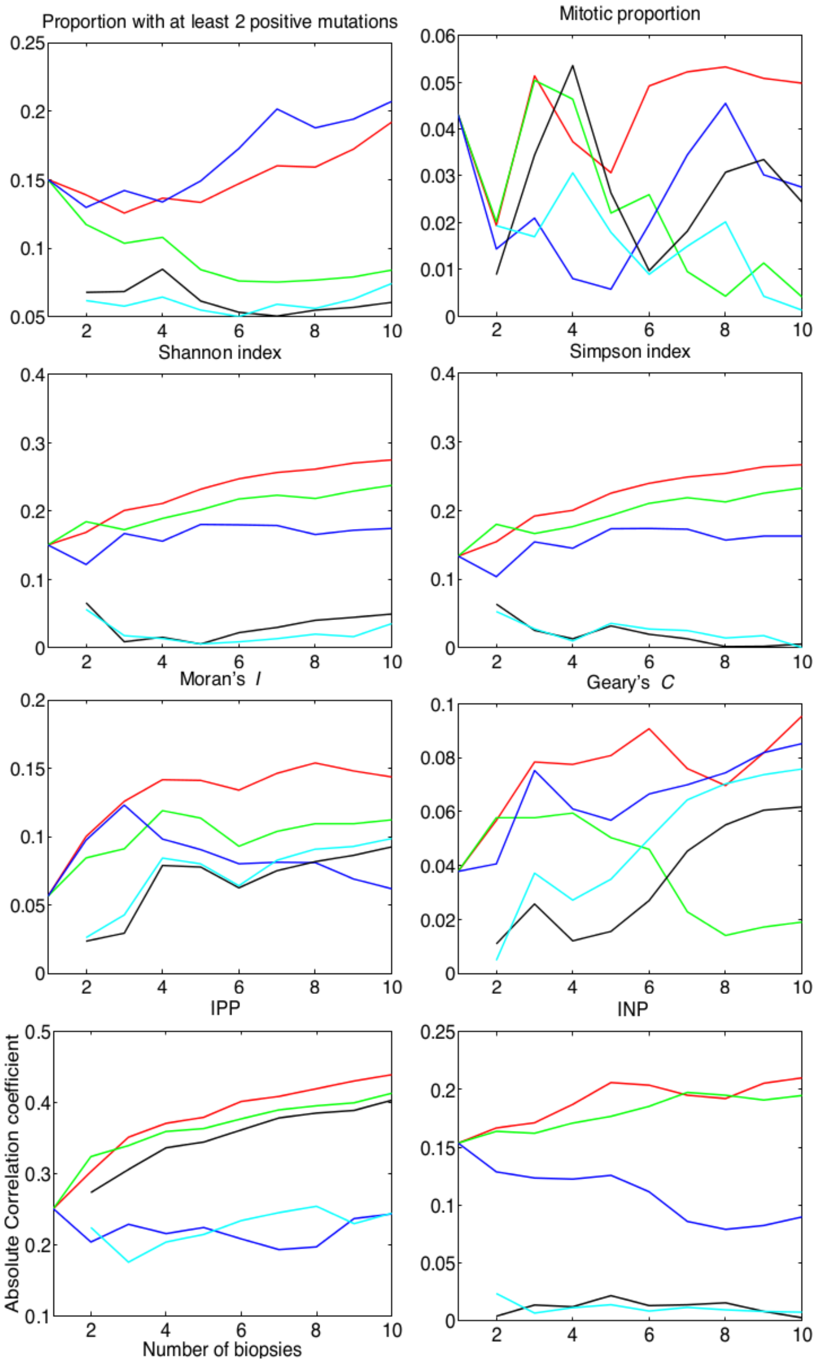


Figure 6. Additional biopsies at the same time point improves prognostication with diminishing returns. Graphs show the relationship between the correlation coefficient (between each biomarker value and waiting time to cancer) and the number of biopsies collected at time $T_b = 50$. Lines denote different measures based on the multiple biopsies: average biomarker value across biopsies (red); maximum value (green); minimum value (blue); difference between maximum and minimum values (black); and variance in values (cyan).

Further, we evaluated a range of different tissue sampling schemes (single biopsy, multiple biopsies in space or time, or random sampling of a lesion). We found that random sampling (such as via an endoscopic brush) provided more consistent prognostic value that a single biopsy, likely because a single (randomly targeted) biopsy is liable to miss localised but 'important' clones. Prognostication was improved by taking multiple biopsies, but with diminishing returns for each additional biopsy taken. Together these data provide a rationale for the empirical evaluation of different tissue sampling schemes.

Interestingly, we found that the difference in biomarker values between two time points is generally not a more prognostic measure than simply using the biomarker value at the later time point alone. This result was counter to our intuition that taking longitudinal biopsies would better characterise the evolutionary trajectory of the lesion and so improve prognostication: in fact we observed that prognostication was improved simply by taking a biopsy from a later time-point. This result illustrates how our *in silico* approach can challenge assumed best-practice and in so doing provide novel insights into biomarker development.

We developed a new statistic, termed the index of positive proliferation (IPP), that proved to be a highly prognostic measure. The IPP is a measure of the average distance to a proliferating cell that has acquired advantageous mutations, and so combines both genetic (or phenotypic) information along with both spatial (cell position) and dynamic (proliferation) information: this integration of multiple different sources of information may somewhat explain the prognostic value of the statistic. empirical measurement of the IPP would be feasible if, for example, the number of driver mutations borne by a cell could be quantified concomitantly with a marker of proliferation (such as Ki-67 staining). Developments in *in situ* genotyping methods might facilitate such an approach in the near future. Irrespective of the immediate feasibility of such a measure, our development and testing of the IPP statistic within our computational model illustrates how *in silico* approaches provide a powerful means to rapidly explore new potential biomarker assays.

Our computational model of cancer evolution is clearly a highly simplified description of reality. For example, we modelled a simple two-dimensional sheet of epithelial cells and neglected the important influence, and indeed co-evolution, of the supporting stroma. We assumed 1:1:1 relationships between genotype, phenotype and fitness, which were further independent of context. Critically, we also used an abstract fitness function to define cellular phenotypes, and in doing so neglected to describe any molecular details of cell behaviour. Adequately describing these kinds of important biological complexities within a model is a necessary step for the development of *in silico* biomarker development platform that is of general use. Increasing the realism of the model would improve confidence that the predicted prognostic value of any biomarker was not an artefact of the over-simplified model, and critically also to facilitate the *in silico* testing of the prognostic value full range of specific biological features (for example the expression of a protein that fufills a particular biological function, such as modulating cell adhesion).

Our study demonstrates how a computational model offers a platform for the initial development of novel prognostic biomarkers: computational can be viewed as a high-throughput and cost-effective screening tool with which to identify the most promising biomarkers for subsequent empirical testing. This work provides the rationale for constructing an *in silico* biomarker development platform that would lessen the current restrictions imposed by the sole reliance on empirical testing.

Acknowledgments

AD is supported by a J.D. Hatcher Award, School of Medicine, Queen's University, Canada. AGF is supported by the EPSRC through grant EP/I017909/1 (www.2020science.net). TAG is supported by Cancer Research UK.

Supporting Information

S1 Software

Zipped folder containing MATLAB implementation of the computational model and analysis described in this study. A thoroughly documented example of how to run the code is provided in the file README.txt.

S1 Text

Supplementary text providing further details on: parameter sensitivity analysis in the full Moran model, variations on proliferative indices, and the impact of choice of local neighbourhood and update rule on the probability and time to fixation in a simplified version of the mathematical model.

In the following sections we examine the effect of model parameter values and choice of simulation method on the results of our analysis of the computational model.

Robustness of results to mutation rate, μ. To determine the effects of the mutation rate, μ, on the predictive value of each biomarker, we considered the two cases $\mu = 0.05$ and $\mu = 0.01$, with all other parameters set to their default values. Hazards ratios are computed for these cases in Tables S1 and S2, respectively. We found that for smaller mutation rates, the predictive value of each biomarker decreases, due to the longer timescale over which cancer occurs. We note in particular that the IPP measure remained predictive in biopsy samples, even for the lowest value of μ considered.

Robustness of results to biopsy size, N_b. To determine the effects of the biopsy size, N_b, on the predictive value of each biomarker, we considered the two cases $N_b = 5$ and $N_b = 40$, with all other parameters set to their default values. Hazards ratios are computed for these cases in Tables S3 and S4, respectively. As expected, when a larger biopsy is taken the predictive value of each biomarker increases. For the small biopsy ($N_b = 5$) we found that no predictive values achieved statistical significance, while for the large biopsy ($N_b = 40$), the predictive value of the Simpson index and IPP measures exhibited statistical significance. These results suggest that there is a critical biopsy size at which *all* biomarker values may gain predictive value and statistical significance, and the sampling effects of noise can be overcome.

Robustness of results to threshold for cancer, N_m. To examine the effects of the number of advantageous mutations necessary for a cell to be considered cancerous, N_m, on the predictive value of each biomarker, we considered the three cases $N_b \in \{3, 5, 15\}$, with all other parameters set to their default values. Hazards ratios are computed for these cases in Tables S5, S6 and S7, respectively. In particular, we found that the IPP measure consistently provided a strong and statistically-significant hazards ratio when computed from a biopsy or whole-lesion sample, and its predictive value increased with N_m.

Robustness of results to fitness changes due to mutation, s_p and s_d. To examine the effects of the degree to which non-neutral mutations alters cellular fitness on the predictive value of each biomarker, we considered the three cases $s_p = s_d \in \{0, 0.002, 0.02\}$, with all other parameters set to their default values. Hazards ratios are computed for these cases in Tables S8, S0 and S10, respectively. In particular, we found that as selective advantage or disadvantage is varied, the IPP measure remained predictive of outcome, especially in biopsy samples, underscoring its robustness. In addition the Simpson index showed strong predictive value when taken on the whole lesion in each case.

Variations on proliferation indices. We next explored further variations on the proliferation indices defined in the main text. First, we varied the number of advantageous mutations required to be considered a ‘positive cell’ by examining the cases defined as IPP 1 and IPP 3, where the

f_i in equation (8) are re-defined such that for IPP k we have

$$f_i^{(k)} = \begin{cases} 0: & m_i < N_m - k, \\ 1: & m_i \geq N_m - k, \end{cases} \tag{10}$$

and the remainder of the formula for the IPP measure is unchanged. In addition, we consider a measure termed the index of mutation proliferation (IMP), defined analogously to the IPP, but with a choice of fitness function such that $f^{(1)}_{\mathrm{IMP},i}$ is the number of advantageous mutations obtained by cell i (IMP 1), and $f^{(2)}_{\mathrm{IMP},i}$ is the total number of mutations (the sum of advantageous, deleterious and neutral mutations) accumulated by cell i (IMP 2).

Results depicting the predictive value of these additional indices, as a function of sampling time, on both the whole-lesion and biopsy samples is summarised in Fig. S1. These results indicate that the behaviour of IPP1 and IPP3 is broadly similar to that of the IPP, albeit with their predictive values peaking at different times.

Effect of cell neighbourhood in a simplified model. Having investigated the sensitivity of our results to the values of model parameters, we next examined whether the choice of cell neighbourhood has any significant effect on model behaviour. In lattice-based models such as our spatial Moran process, two common choices for defining the local topology are the Moore and von Neumann neighbourhoods. The von Neumann neighbours of a given cell are those that lie one lattice spacing from it by the Manhattan metric, while the Moore neighbours are those that lie less than two lattice spacings from it by the Euclidean metric.

To investigate whether the choice of neighbourhood has any impact on the behaviour of our mathematical model, we considered a simplified model with two cell types, which we refer to as normal and mutant. We employed the 'death-birth' update rule as defined in the main text. For simplicity we considered an irreversible neutral mutation giving no relative fitness advantage to mutants, with two cases of mutation probability considered: $\mu = 0$ (corresponding to a classical spatial Moran process), and $\mu = 0.3$. We considered a small 10×10 lattice comprised of cells and introduced a single mutant cell at the centre of the lattice. The model was simulated until the lattice was composed entirely of non-mutant cells or entirely of mutant cells (one of these possibilities must occur eventually, since they are both absorbing states of the model).

The cumulative distribution function for the time to fixation for the mutant cell population, computed based on 10^5 runs of the above model using either Moore or von Neumann cell neighbourhoods, is shown in Fig. S2. We found that as the mutation rate is increased, the difference in the CDFs becomes nearly negligible. Additionally, in the case where $\mu = 0$, for the von Neumann neighbourhood case, we found that 1.13% of the 10^5 runs end in mutant fixation, and in the case of the Moore neighbourhood, 1.19% of the runs end in mutant fixation. When

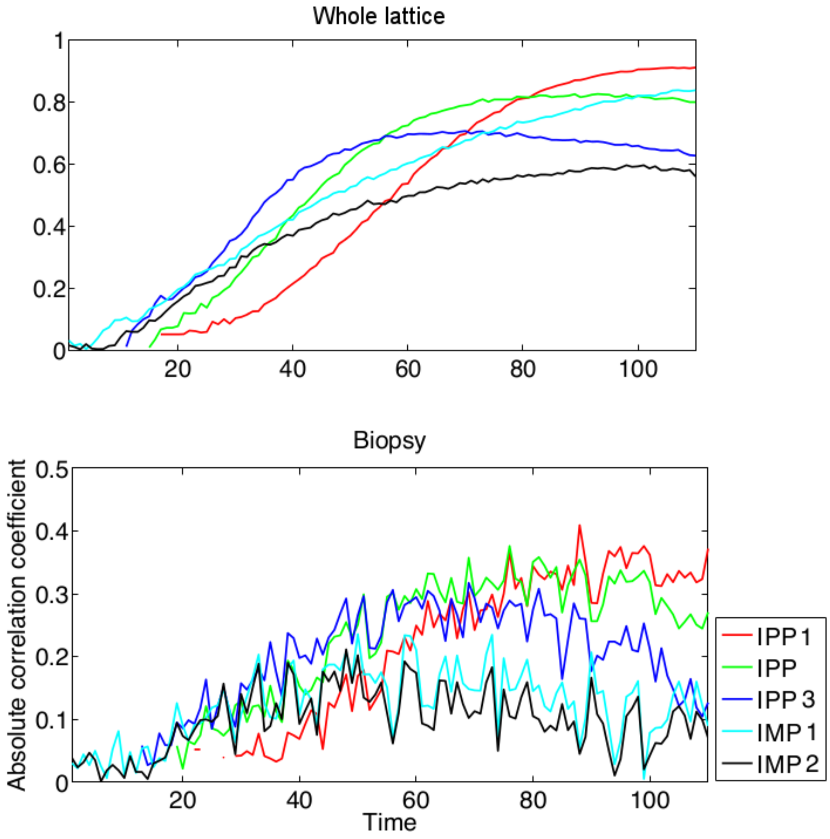
μ = 0.3, the probability of mutant fixation is higher, and for the von Neumann neighbourhood case, we find that 97.7% of the 2000 runs end in mutant fixation, whereas for the Moore neighbourhood the proportion is 98.2%. In both cases, the differences between the two cases of neighbourhood types diminished as μ increased, and limiting-time behaviour cases for both systems was highly similar.

Comparison of model update rules in a simplified model. Finally, to examine the effect that the choice of update rule has on the behaviour of our model, we considered 8 different update rules and compared key summary statistics in each case, for the simplified model described in the previous section. In this case, we incorporated an advantage to the mutations incurred by the mutant cells, conferring upon them a relative fitness of $s_p = 1.2$.

We use a two-letter notation to described the different choices of update rule, where the first and second letters represent the choice of first and second steps in the update rule, respectively. We use the letters *b* and *d* to denote birth (division) and death (removal), and we use a capital letter to denote the case where the corresponding step is influenced by cellular fitness, enabling the incorporation of selection. For instance, *bD* represents the update rule where first a cell is chosen uniformly at random to divide, and then a neighbouring cell is chosen to die, biased by inverse fitness.

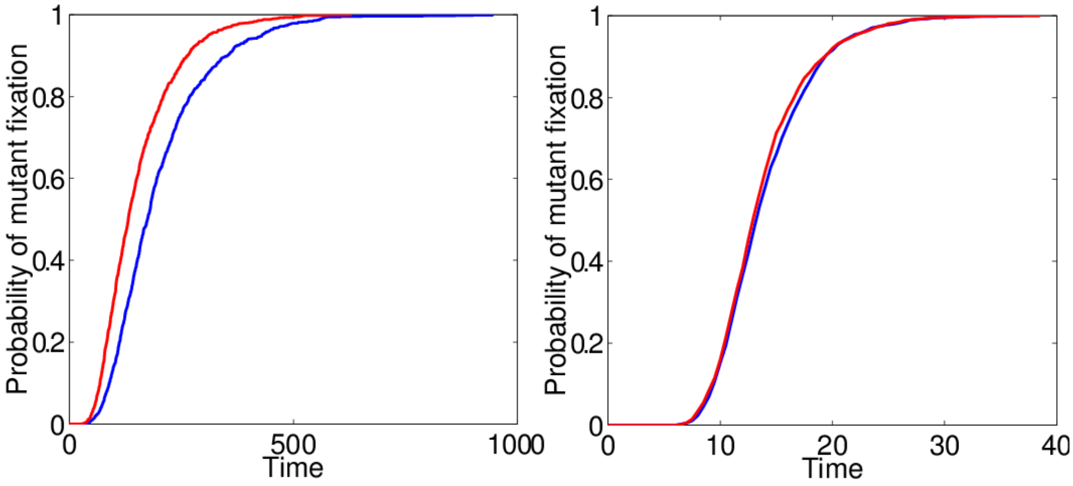
We examine the average time to fixation for a mutant-only population, and the variance in these times to fixation, as well as the cumulative distribution function (CDF) for the time to fixation of the mutant population, conditional on this occurring before the mutants entirely die out. These results are presented for the case where μ = 0 in Table S11 and Fig. S3. We conclude that in this case, the various update rules produce differing results, but that as μ increases, these differences decrease in magnitude, as the times to fixation are reduced, and the probability of mutant fixation approaches unity. That is, the results presented in the above analysis depict upper bounds on the differences, and we note that based on our analysis, these differences decrease as μ is increased above zero, into the regimes of values considered within the present study, and those that are biologically relevant.

S1 Fig



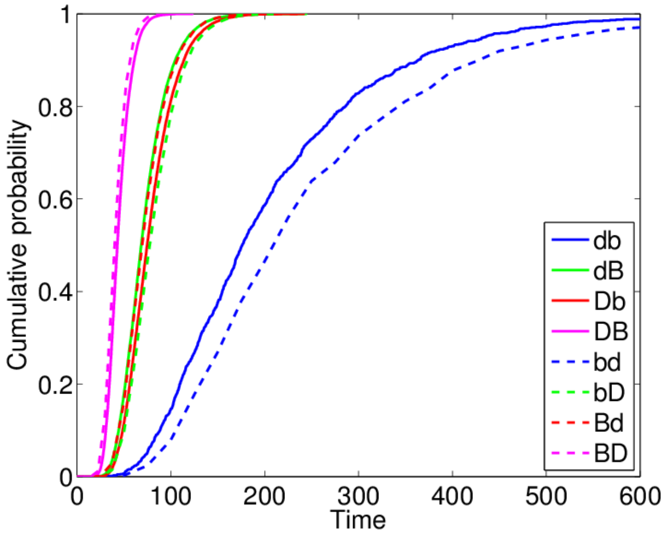
Variation of the absolute value of the correlation coefficient, as a function of sampling time T_b, for additional indices considered in Text S1. Data shown are for the case where $N_m = 10$, $s_p = s_d = 0.2$, and $\mu = 0.1$, over 1000 runs, for indices taken on circular biopsies with radius $N_b = 20$.

S2 Fig



Cumulative distribution functions for the time to mutant fixation in a spatial Moran model with von Neumann and Moore neighbourhoods. The model includes a single, neutral, irreversible mutation ($s_p = s_d = 0$, $\mu = 0$) and each simulation is initiated with a single mutant cell in the centre of the lattice. The blue and red curves correspond to the von Neumann and Moore neighbourhoods, respectively. Left: The case $\mu = 0$; results from 10^5 simulations. Right: The case $\mu = 0.3$; results from 2000 simulations.

S3 Fig



Cumulative distribution functions for the time to mutant fixation in a spatial Moran model under different update rules. The model includes a single, neutral, irreversible mutation ($s_p = s_d = 0$, $\mu = 0$) and each simulation is initiated with a single mutant cell in the centre of the lattice. Legend: B and D correspond to birth and death respectively, written in the order in which these processes are implemented in the update rule, with capitalized letters indicating random selection of cells biased by their inverse fitness to account for the effects of selective advantage. Results are generated from 2000 simulations of a small 10 × 10 lattice.

S1 Table

Summary of Cox proportional hazards models for the case of a lower mutation rate ($\mu = 0.005$). Hazard ratios (HR150) are computed at time $t = 150$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $N_m = 10$, $s_p = s_d = 0.02$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR150	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	1	$(0, \infty)$	1
	Mitotic proportion	0.01	0.23	(0.023, 2.3)	0.21
	Shannon index	0.01	**1.1**	(1, 1.1)	$< 10^{-4}$
	Simpson index	0.01	**3.1**	(2.2, 4.3)	$< 10^{-4}$
	Moran's I	0.01	**1.2**	(1.1, 1.3)	0.00096
	Geary's C	0.01	0.96	(0.89, 1)	0.38
	IPP	0.01	**1.4**	(1.3, 1.4)	$< 10^{-4}$
	INP	0.01	1.8	(0.76, 4.5)	0.18
Biopsy	$n_p > 1$ proportion	0.01	1	(0.93, 1.1)	0.83
	Mitotic proportion	0.01	1	(0.41, 2.4)	1
	Shannon index	0.01	1	(1,1)	0.074
	Simpson index	0.01	1.7	(0.93, 3)	0.085
	Moran's I	0.01	**1.1**	(1, 1.1)	0.0034
	Geary's C	0.01	1	(0.99, 1)	0.33
	IPP	0.1	**1.5**	(1.3, 1.7)	$< 10^{-4}$
	INP	0.01	0.99	(0.98,1)	0.21
Scraping	$n_p > 1$ proportion	0.01	1	$(0, \infty)$	1
	Mitotic proportion	0.01	0.84	(0.33, 2.2)	0.72
	Shannon index	0.05	**1.3**	(1.2, 1.4)	$< 10^{-4}$
	Simpson index	0.01	**2.4**	(1.7, 3.2)	$< 10^{-4}$

S2 Table

Summary of Cox proportional hazards models for the case of a much lower mutation rate (μ = 0.001). Hazard ratios (HR375) are computed at time t = 375 for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are N_b = 20, N_m = 10, $s_p = s_d$ = 0.02 and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR375	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	1	$(0, \infty)$	1
	Mitotic proportion	0.01	0.24	$(0.012, 4.8)$	0.35
	Shannon index	0.1	1.1	$(1, 1.2)$	0.027
	Simpson index	0.01	**1.1**	$(1, 1.1)$	0.027
	Moran's I	0.01	1	$(0.96, 1.1)$	0.58
	Geary's C	0.01	0.97	$(0.93, 1)$	0.24
	IPP	0.01	**1.1**	$(1.1, 1.1)$	$< 10^{-4}$
	INP	0.01	13	$(0.049, \infty)$	0.37
Biopsy	$n_p > 1$ proportion	0.01	1.2	$(0.97, 1.5)$	0.089
	Mitotic proportion	0.01	0.78	$(0.33, 1.8)$	0.57
	Shannon index	0.01	1	$(1, 1)$	0.36
	Simpson index	0.01	1.1	$(0.93, 1.2)$	0.34
	Moran's I	0.01	1	$(0.98, 1)$	0.44
	Geary's C	0.01	1	$(0.98, 1)$	0.9
	IPP	0.1	1.2	$(1.1, 1.2)$	$< 10^{-4}$
	INP	0.01	1	$(0.99, 1)$	0.37
Scraping	$n_p > 1$ proportion	0.01	1	$(0, \infty)$	1
	Mitotic proportion	0.01	0.77	$(0.36, 1.7)$	0.51
	Shannon index	0.01	**1**	$(1, 1)$	0.026
	Simpson index	0.01	**1.1**	$(1, 1.1)$	0.024

S3 Table

Summary of Cox proportional hazards models for the case of a smaller biopsy ($N_b = 5$). Hazard ratios (HR75) are computed at time $t = 75$ for various putative biomarkers, for a circular biopsy. Other parameter values used are $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR75	95% CI	p
Biopsy	$n_p > 1$ proportion	0.01	1	(0.98, 1)	0.75
	Mitotic proportion	0.01	1	(0.87, 1.2)	0.67
	Shannon index	0.01	1	(0.99, 1)	0.27
	Simpson index	0.01	0.94	(0.84, 1.1)	0.3
	Moran's I	0.01	1	(0.96, 1)	0.93
	Geary's C	0.01	1	(0.99, 1)	0.44
	IPP	0.01	1	(0.97, 1)	0.81
	INP	0.01	1	(1, 1)	0.95

S4 Table

Summary of Cox proportional hazards models for the case of a larger biopsy ($N_b = 40$). Hazard ratios (HR12) are computed at time $t = 12$ for various putative biomarker schemes, for a circular biopsy. Other parameter values used are $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR75	95% CI	p
Biopsy	$n_p > 1$ proportion	0.01	1	(0.92, 1.2)	0.6
	Mitotic proportion	0.01	2.1	(0.46, 9.6)	0.34
	Shannon index	0.1	1.2	(1, 1.5)	0.018
	Simpson index	0.01	**11**	(1.5, 87)	0.018
	Moran's *I*	0.01	1	(0.9, 1.2)	0.67
	Geary's *C*	0.01	1	(0.97, 1)	0.82
	IPP	0.01	**1.3**	(1.2, 1.4)	$< 10^{-4}$
	INP	0.01	1.1	(0.96, 1.2)	0.19

S5 Table

Summary of Cox proportional hazards models for the case of a much lower threshold number of mutations for cancer (N_m = 3). Hazard ratios (HR12) are computed at time $t = 12$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $s_p = s_d = 0.2$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR12	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	**2.4**	(2.1, 2.8)	$< 10^{-4}$
	Mitotic proportion	0.01	1.3	(0.27, 6.3)	0.73
	Shannon index	0.01	**1.2**	(1.1, 1.2)	$< 10^{-4}$
	Simpson index	0.01	**2.7**	(2, 3.7)	$< 10^{-4}$
	Moran's I	0.01	**6.6**	(2.1, 20)	0.0011
	Geary's C	0.01	1	(0.8, 1.3)	0.9
	IPP	0.01	**1.3**	(1.2, 1.5)	$< 10^{-4}$
	INP	0.01	**1.4**	(1.2, 1.5)	$< 10^{-4}$
Biopsy	$n_p > 1$ proportion	0.01	1.1	(0.87, 1.5)	0.37
	Mitotic proportion	0.01	0.73	(0.43, 1.2)	0.25
	Shannon index	0.01	1	(0.99, 1)	0.62
	Simpson index	0.01	1.1	(0.81, 1.6)	0.49
	Moran's I	0.01	1.1	(0.88, 1.4)	0.4
	Geary's C	0.01	1	(0.98, 1)	0.74
	IPP	0.1	1.6	(1.2, 2.1)	0.0028
	INP	0.1	1.6	(1.1, 2.3)	0.0083
Scraping	$n_p > 1$ proportion	0.01	**1.4**	(1.3, 1.5)	$< 10^{-4}$
	Mitotic proportion	0.01	1.4	(0.92, 2.2)	0.11
	Shannon index	0.1	1.5	(1.1, 2.1)	0.0049
	Simpson index	0.01	**1.3**	(1, 1.5)	0.015

S6 Table

Summary of Cox proportional hazards models for the case of a lower threshold number of mutations for cancer ($N_m = 5$). Hazard ratios (HR24) are computed at time $t = 24$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $s_p = s_d = 0.2$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR24	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	**1.2**	(1.1, 1.3)	$< 10^{-4}$
	Mitotic proportion	0.01	0.46	(0.067, 3.2)	0.44
	Shannon index	0.01	**1.1**	(1.1, 1.2)	$< 10^{-4}$
	Simpson index	0.01	**3.8**	(2.5, 5.7)	$< 10^{-4}$
	Moran's *I*	0.01	1.6	(0.87, 2.8)	0.13
	Geary's *C*	0.01	0.9	(0.76, 1.1)	0.23
	IPP	0.01	**1.5**	(1.4, 1.7)	$< 10^{-4}$
	INP	0.01	**1.3**	(1.2, 1.4)	$< 10^{-4}$
Biopsy	$n_p > 1$ proportion	0.01	1.1	(0.94, 1.2)	0.28
	Mitotic proportion	0.01	0.89	(0.49, 1.6)	0.71
	Shannon index	0.01	1	(1, 1)	0.082
	Simpson index	0.01	1.8	(0.96, 3.3)	0.069
	Moran's *I*	0.01	1.1	(0.99, 1.3)	0.076
	Geary's *C*	0.01	0.99	(0.96, 1)	0.41
	IPP	0.1	1.5	(1.2, 2)	0.00069
	INP	0.01	**1.1**	(1, 1.1)	0.00011
Scraping	$n_p > 1$ proportion	0.01	**1.1**	(1.1, 1.2)	$< 10^{-4}$
	Mitotic proportion	0.01	0.6	(0.34, 1.1)	0.075
	Shannon index	0.01	**1.1**	(1, 1.1)	$< 10^{-4}$
	Simpson index	0.01	**2.3**	(1.7, 3.2)	$< 10^{-4}$

S7 Table

Summary of Cox proportional hazards models for the case of a larger threshold number of mutations for cancer (N_m = 15). Hazard ratios (HR225) are computed at time t = 225 for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are N_b = 20, s_p = s_d = 0.2, μ = 0.1 and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR225	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	1	(0, ∞)	1
	Mitotic proportion	0.01	**0.015**	(0, 0.51)	0.019
	Shannon index	0.01	**1.1**	(1.1, 1.1)	$< 10^{-4}$
	Simpson index	0.01	**13**	(5.8, 29)	$< 10^{-4}$
	Moran's *I*	0.01	**1.3**	(1.1, 1.5)	0.00079
	Geary's *C*	0.01	0.91	(0.83, 1)	0.051
	IPP	0.01	**1.7**	(1.6, 1.8)	$< 10^{-4}$
	INP	0.01	0.0059	(0, ∞)	0.93
Biopsy	$n_p > 1$ proportion	0.01	1	(0.96, 1)	0.77
	Mitotic proportion	0.01	1.1	(0.44, 2.8)	0.83
	Shannon index	0.01	1	(1, 1)	0.067
	Simpson index	0.01	3.2	(1, 11)	0.051
	Moran's *I*	0.01	0.99	(0.95, 1)	0.58
	Geary's *C*	0.01	0.99	(0.98, 1)	0.57
	IPP	0.1	1.4	(1.2, 1.8)	0.00013
	INP	0.01	1	(1, 1)	0.67
Scraping	$n_p > 1$ proportion	0.01	1	(0, ∞)	1
	Mitotic proportion	0.01	0.64	(0.23, 1.8)	0.4
	Shannon index	0.01	**1.1**	(1,1.1)	$< 10^{-4}$
	Simpson index	0.01	**6.7**	(3.4, 13)	$< 10^{-4}$

S8 Table

Summary of Cox proportional hazards models for the case where mutations do not alter cellular fitness ($s_p = s_d = 0$). Hazard ratios (HR105) are computed at time $t = 105$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $N_m = 10$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR105	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	**1.1**	$(1, 1.2)$	0.0022
	Mitotic proportion	0.01	1.2	$(0.26, 5.2)$	0.85
	Shannon index	0.01	**1.1**	$(1, 1.1)$	$< 10^{-4}$
	Simpson index	0.01	**570**	$(6.7, \infty)$	0.0052
	Moran's I	0.01	0.83	$(0.64, 1.1)$	0.16
	Geary's C	0.01	1.1	$(0.97, 1.2)$	0.15
	IPP	0.01	**3.5**	$(2.8, 4.4)$	$< 10^{-4}$
	INP	0.01	**1.1**	$(1, 1.1)$	0.014
Biopsy	$n_p > 1$ proportion	0.01	1	$(0.99, 1)$	0.55
	Mitotic proportion	0.01	0.99	$(0.63, 1.6)$	0.98
	Shannon index	0.01	1.4	$(1.2, 1.8)$	0.054
	Simpson index	0.01	15	$(0.58, 400)$	0.1
	Moran's I	0.01	1	$(0.92, 1.1)$	1
	Geary's C	0.01	1	$(0.99, 1)$	0.34
	IPP	0.01	**1.1**	$(1.1, 1.2)$	$< 10^{-4}$
	INP	0.01	1	$(0.99, 1)$	0.16
Scraping	$n_p > 1$ proportion	0.01	**1.1**	$(1.1, 1.2)$	0.0003
	Mitotic proportion	0.01	1	$(0.67, 1.5)$	0.98
	Shannon index	0.1	1.5	$(1.2, 1.9)$	0.0019
	Simpson index	0.01	**130**	$(3.5, \infty)$	0.009

S9 Table

Summary of Cox proportional hazards models for the case where mutations alter cellular fitness to a much lower extent ($s_p = s_d = 0.0002$). Hazard ratios (HR105) are computed at time $t = 105$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $N_m = 10$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR105	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	1.1	$(0.99, 1.1)$	0.076
	Mitotic proportion	0.01	0.55	$(0.13, 2.2)$	0.4
	Shannon index	0.01	**1.1**	$(1.1, 1.1)$	$< 10^{-4}$
	Simpson index	0.01	**6.6×10^5**	(∞, ∞)	$< 10^{-4}$
	Moran's I	0.01	**1.5**	$(1.1, 2)$	0.0049
	Geary's C	0.01	0.97	$(0.85, 1.1)$	0.63
	IPP	0.01	**2.9**	$(2.5, 3.4)$	$< 10^{-4}$
	INP	0.01	**1.1**	$(1, 1.2)$	0.0029
Biopsy	$n_p > 1$ proportion	0.01	1	$(0.99, 1)$	0.65
	Mitotic proportion	0.01	0.86	$(0.51, 1.5)$	0.58
	Shannon index	0.01	1	$(1, 1.1)$	0.00042
	Simpson index	0.01	**360**	$(12, \infty)$	0.00067
	Moran's I	0.01	1	$(0.95, 1.1)$	0.6
	Geary's C	0.01	0.99	$(0.98, 1)$	0.52
	IPP	0.01	**1.1**	$(1, 1.2)$	0.0018
	INP	0.01	1	$(0.99, 1)$	0.21
Scraping	$n_p > 1$ proportion	0.01	1	$(0.98, 1.1)$	0.17
	Mitotic proportion	0.01	0.67	$(0.44, 1)$	0.065
	Shannon index	0.01	**1.1**	$(1, 1.1)$	$< 10^{-4}$
	Simpson index	0.01	**8500**	$(190, \infty)$	$< 10^{-4}$

S10 Table

Summary of Cox proportional hazards models for the case where mutations alter cellular fitness to a lower extent ($s_p = s_d = 0.002$). Hazard ratios (HR75) are computed at time $t = 75$ for various putative biomarker schemes, for different tissue sampling schemes. Other parameter values used are $N_b = 20$, $N_m = 10$, $\mu = 0.1$ and those listed in Table 1. Statistically significant values are in bold.

		Unit change	HR75	95% CI	p
Whole lesion	$n_p > 1$ proportion	0.01	1.1	$(0.99, 1.2)$	0.078
	Mitotic proportion	0.01	0.75	$(0.18, 3.2)$	0.7
	Shannon index	0.01	**1.1**	$(1.1, 1.1)$	$< 10^{-4}$
	Simpson index	0.01	**6300**	$(330, \infty)$	$< 10^{-4}$
	Moran's I	0.01	1	$(0.73, 1.5)$	0.83
	Geary's C	0.01	0.98	$(0.86, 1.1)$	0.76
	IPP	0.01	**4.3**	$(3.4, 5.5)$	$< 10^{-4}$
	INP	0.01	**1.1**	$(1, 1.2)$	0.00084
Biopsy	$n_p > 1$ proportion	0.01	1	$(0.99, 1)$	0.48
	Mitotic proportion	0.01	1.1	$(0.69, 1.8)$	0.65
	Shannon index	0.01	1	$(0.99, 1)$	0.3
	Simpson index	0.01	3.4	$(0.22, 53)$	0.38
	Moran's I	0.01	1	$(0.94, 1.1)$	0.66
	Geary's C	0.01	1	$(0.99, 1)$	0.61
	IPP	0.01	**1.1**	$(1, 1.2)$	0.0017
	INP	0.01	1	$(0.99, 1)$	0.47
Scraping	$n_p > 1$ proportion	0.01	1	$(0.98, 1.1)$	0.19
	Mitotic proportion	0.01	1.4	$(0.89, 2.3)$	0.14
	Shannon index	0.01	**1.1**	$(1.1, 1.1)$	$< 10^{-4}$
	Simpson index	0.01	**1100**	$(91, \infty)$	$< 10^{-4}$

S11 Table

Effect of update rule on mutant fixation statistics in a spatial Moran model with selection only. Summarized statistics from 10^5 runs of each update rule used in the simplified Moran model described in this section, in the case $\mu = 0$. Proportion: the proportion of simulations in which mutant fixation occurred. Mean, variance: the mean and variance of the times to mutant fixation, across simulations in which this occurred. Simulations were run until the population consisted entirely of either non-mutant cells or mutant cells.

Update rule	Proportion	Mean	Variance
db	0.011	206	14144
dB	0.162	72.3	614
Db	0.170	78.7	700
DB	0.296	44.8	146
bd	0.009	244	19390
bD	0.154	81.3	719
Bd	0.162	73	595
BD	0.294	41.4	121

References

1. Reid BJ, Kostadinov R, Maley CC. New strategies in Barrett's esophagus: integrating clonal evolutionary theory with clinical management. Clin Cancer Res. 2011;17:3512–3519.
2. Jones JL. Progression of ductal carcinoma in situ: the pathological perspective. Breast Cancer Res. 2006;8:204.
3. Crawford ED. Understanding the epidemiology, natural history, and key pathways involved in prostate cancer. Urology. 2009;73:S4–S10.
4. Miyamoto H, Miller JS, Fajardo DA, Lee TK, Netto GJ, Epstein JI. Non-invasive papillary urothelial neoplasms: The 2004 WHO/ISUP classification system. Pathol Int. 2010;60:1–8.
5. Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. New Engl J Med. 2011;365:1375–1383.
6. Coldiron BM, Mellette JR, Hruza GJ, Helm TN, Garcia CA. Addressing overdiagnosis and overtreatment in cancer. Lancet Oncol. 2014;15:e307.
7. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006;24:3726–3734.
8. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304:1497–1500.
9. Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Reid BJ. Selectively advantageous mutations and hitchhikers in neoplasms p16 lesions are selected in Barrett's Esophagus. Cancer Res. 2004;64(10):3414–3427.
10. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer. 2005;5:845–856.
11. Lari SA, Kuerer HM. Biological markers in DCIS and risk of breast recurrence: a systematic review. J Cancer. 2011;2:232–261.
12. Thorsteinsdottir S, Gudjonsson T, Nielsen OH, Vainer B, Seidelin JB. Pathogenesis and biomarkers of carcinogenesis in ulcerative colitis. Nat Rev Gastroenterol Hepatol. 2011;8:395–404.
13. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12:323–334.
14. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA. 2006;103:5923–5928.

15. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet. 2006;38(4):468–473.
16. Park SY, Gönen M, Kim HJ, Michor F, Polyak K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. J Clin Invest. 2010;120(2):636.
17. Bochtler T, Stölzel F, Heilig CE, Kunz C, Mohr B, Jauch A, et al. Clonal heterogeneity as detected by metaphase karyotyping is an indicator of poor prognosis in acute myeloid leukemia. J Clin Oncol. 2013;p. JCO–2013.
18. Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C. Intratumor heterogeneity: seeing the wood for the trees. Sci Transl Med. 2012;4(127):127ps10.
19. Almendro V, Cheng YK, Randles A, Itzkovitz S, Marusyk A, Ametller E, et al. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. Cell Rep. 2014;6(3):514–527.
20. Michor F, Iwasa Y, Nowak MA. Dynamics of cancer progression. Nat Rev Cancer. 2004;4:197–205.
21. Durrett R, Moseley S. Evolution of resistance and progression to disease during clonal expansion of cancer. Theor Pop Biol. 2010;77:42–48.
22. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, et al. Genetic progression and the waiting time to cancer. PLoS Comput Biol. 2007;3:e225.
23. Bozic I, Antal Tr, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci USA. 2010;107:18545–18550.
24. Martens EA, Kostadinov R, Maley CC, Hallatschek O. Spatial structure increases the waiting time for cancer. New J Phys. 2011;13:115014.
25. Anderson ARA, Weaver AM, Cummings PT, Quaranta V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. Cell. 2006;127:905–915.
26. Anderson ARA, Hassanein M, Branch KM, Lu J, Lobdell NA, Maier J, et al. Microenvironmental independence associated with tumor progression. Cancer Res. 2009;69:8797–8806.
27. Korolev KS, Xavier JB, Gore J. Turning ecology and evolution against cancer. Nat Rev Cancer. 2014;14:371–380.
28. Williams T, Bjerknes R. Stochastic model for abnormal clone spread through epithelial basal layer. Nature. 1972;236:19–21.
29. Foo J, Leder K, Ryser MD. Multifocality and recurrence risk: a quantitative model of field

cancerization. J Theor Biol. 2014;355:170–184.

30. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys. 1976;22:403–434.
31. Kuukasjärvi T, Kononen J, Helin H, Holli K, Isola J. Loss of estrogen receptor in recurrent breast cancer is associated with poor response to endocrine therapy. J Clin Oncol. 1996;14:2584–2589.
32. Kröger N, Milde-Langosch K, Riethdorf S, Schmoor C, Schumacher M, Zander AR, et al. Prognostic and predictive effects of immunohistochemical factors in high-risk primary breast cancer patients. Clin Cancer Res. 2006;12:159–168.
33. Arteaga CL, Sliwkowski MX, Osborne CK, Perez EA, Puglisi F, Gianni L. Treatment of HER2-positive breast cancer: current status and future perspectives. Nat Rev Clin Oncol. 2011;9:16–32.
34. Shannon CE. Communication theory of secrecy systems. Bell Syst Tech J. 1949;28:656–715.
35. Simpson EH. Measurement of diversity. Nature. 1949;163:688–688.
36. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37:17–23.
37. Geary RC. The contiguity ratio and statistical mapping. The Incorporated Statistician. 1954;5:115–146.
38. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New Engl J Med. 2012;366:883–892.