
Article:
Dhawan, A, Graham, T and Fletcher, AG (2016) A computational modelling approach for deriving biomarkers to predict cancer risk in premalignant disease. Cancer Prevention Research . ISSN 1940-6207

https://doi.org/10.1158/1940-6207.CAPR-15-0248

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

Andrew Dhawan[1], Trevor A. Graham[2,*], Alexander G. Fletcher[3,4,*]

[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
[4] School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

Running title: Biomarker evaluation for premalignant disease *in silico*

Keywords: Premalignant lesions; biomarkers; screening; simulation; evolution.

Financial support: A. Dhawan acknowledges support from a J.D. Hatcher Award, School of Medicine, Queen's University, Canada. T.A. Graham acknowledges support from Cancer Research UK. A.G. Fletcher acknowledges support by the UK Engineering and Physical Sciences Research Council through grant EP/I017909/1 (www.2020science.net).

*** Corresponding authors:**
Alexander G. Fletcher
School of Mathematics and Statistics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield, S3 7RH, UK
Tel: +44 (0)114 222 3846
Email: a.g.fletcher@sheffield.ac.uk

Trevor A. Graham
Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK
Tel: +44 (0)20 7882 6231
Email: t.graham@qmul.ac.uk

Conflict of interest statement: The authors declare no conflict of interest.

Word count: 6000
Abstract word count: 219
Number of tables: 2
Number of figures: 7
Number of supplementary data: 24 (17 supplementary tables, 5 supplementary figures, 1 document of supplementary figure legends, 1 document of supplementary text)

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 slightly 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 morphological 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 premalignant 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 [15–17]. 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.

Materials and 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 emphasize that a cell that has accumulated mutations behaves the same as a wild-type cell in terms of mode of division and accumulation of mutations; the only difference between cells lies in their relative fitness, and hence the probability that they are chosen for removal as specified by the death-birth process.

We define the *time of 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 time of clinically detectable 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 Text 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 amplification staining commonly performed for malignancies of the breast [31–33]. Such assays are cost-effective and relatively simple to implement.

Here, we use the cutoff of a cell having acquired at least N_p 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 N_p advantageous mutations, divided by the total number of cells sampled. We present results based on $N_p = 2$ throughout the text, thus using the shorthand $N_p > 1$ to refer to this biomarker. We discuss the robustness of these results to the chosen value of $N_p = 2$ in the Results section.

Mitotic proportion. Proliferative cells are usually identified in tissue sections or cytology specimens using immunohistochemistry for cell-cycle associated proteins, foremost Ki-67 [34]. These proteins have a natural half-life over which a proliferative cell can be identified. To represent this measure in our computational model, we defined a time window t_w over which a proliferative marker can be detected by staining. The mitotic proportion at a given time t is then defined as the number of cells that have undergone mitosis at least once in the time interval $(t - t_w, t]$, 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 [35]. 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.

Gini-Simpson index. Another established measure of diversity is the Simpson index [36]. To ensure that a higher value corresponds to greater diversity, we choose to use a transformation of this index called the Gini-Simpson index, S, which is defined as follows [37]. For a population of K distinct types, each comprising proportion p_k of the population, we have

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

This index may be thought of as the probability that two randomly chosen members from the population are of different types. As the index decreases towards the minimum of 0, 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 [38]. Moran's *I* takes values in $[-1, 1]$. For a given set of values $\{X_1 \dots X_N\}$, 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. 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 $i, j \in \{1,\dots,N\}$. 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 [39]. 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: we might expect the concentration of proliferation in regions of high numbers of advantageous mutations to correlate with a poor prognosis.

We define the IPP as follows. Consider a population of N cells, with the individual cells labelled as $X_1, \ldots, X_N$. Suppose that a subset of these cells, $Y_1, \ldots, Y_Q$, are proliferating at a given time. We define *cellular contributions* $f_1, \ldots, f_M \in \mathbb{R}^+$ as values such that a higher contribution corresponds to a cellular state genetically closer to that of cancer. Clinically, these cellular contributions 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. For each cell i with m_i advantageous mutations, we define the cellular contribution f_i as

$$f_i = \begin{cases} 0: & m_i < N_m - 2, \\ 1: & m_i \geq N_m - 2, \end{cases} \tag{7}$$

Thus, for a given spatial weight matrix $(w_{ij}) \in \mathbb{R}^{N \times N}$ we define the IPP as

$$\mathrm{IPP} = \frac{\sum_{i=1}^{Q} \sum_{j=1}^{N} w_{ij} f_j}{\sum_{i=1}^{Q} \sum_{j=1}^{N} w_{ij}}. \tag{8}$$

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 $w_{ij} = 0$.

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

contributions 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. The cutoff value of $N_m/2$ was chosen in an *ad hoc* manner based on preliminary simulations; we note that refinement of this parameter may be necessary for effective use of the INP in future studies.

Statistical methods

The statistical association (correlation with the time of clinically 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. For all presented *p*-values, the significance cutoff is taken as 0.05.

Data used for the Cox regression model were all generated by the stochastic simulations of the computational model. That is, the event times were defined as the simulation times at which 5% of the cells of the lattice were defined as cancerous, and the predictors of this time were taken to be the biomarker index values computed at an earlier simulation time. The cohort size is therefore the number of such simulations which were carried out, which was 10^3. There was no censoring required, as all simulations were run to completion of endpoint as defined previously.

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. Time is treated as a continuous variable in the model, but is simulated as a succession of discrete time steps, where the length of each time step is a function of the overall fitness of the population and a stochastic factor, as per the 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 time is defined as the time of clinically detectable cancer. 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. 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 time of clinically detectable cancer (see *Methods*). The model exhibits successive clonal sweeps of mutations (Fig. 1C).

Assessment of candidate biomarkers and tissue sampling schemes

Counting driver mutations. We considered the correlation between the proportion of cells bearing at least N_p advantageous mutations (so-called driver mutations) and the time of clinically detectable cancer. The closer the cutoff N_p is to the number of mutations required for cancer, N_m, the more correlated this measure became with the time of clinically detectable cancer (Table S1). These results confirm the intuition that it is easier to predict the occurrence of a cancer at a time point close to when the cancer will occur (e.g. at the 'end' of the evolutionary process, when only a few additional driver mutations are required) than early in the cancer's development, when many additional mutations are required.

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 Gini-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.

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 ($p = 0.23$; Fig. 2A), 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 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 an early time of clinically detectable cancer ($p < 10^{-4}$ for both the

Shannon index (Fig. 2B) and Gini-Simpson index (Fig. 2C).

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 predictive value of our putative biomarkers. In this case, the proportion of cells with at least one advantageous mutation remained a poor prognosticator ($p = 0.29$), whereas the proportion of proliferative cells became a significant predictor ($p = 0.02$) (see Table 2).
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-correlated.

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

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.

To assess the predictability of each putative biomarker [40] we calculated the area under the receiver operating characteristic (ROC) curves (Fig. 4) as a function of the censoring time [41]. ROC curves are the curves defined by the sensitivity and specificity of each index value as it predicts the 'end time' (time of clinically detectable cancer), where a positive end time is a time past a certain pre-defined simulation time, and the cutoff for the index value that defines whether the index predicts if that end time is early or late, is continuously varied. The area under these curves is 1 in the case of an index value that is perfectly predictive of an end time, and 0.5 for random guessing as to whether the index predicts if the end time is early or late. These curves show the IPP measure has the best predictive value of all measures considered, and that the Shannon and Gini-Simpson diversity indices also have strong predictive value. The lack of predictive value derived from the mitotic proportion, Geary's C and proportion of mutant cells was also confirmed.

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 biopsy affects its prognostic value by investigating how the correlation coefficient between each biomarker and the subsequent time of clinically detectable 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 statistical association than biopsies collected earlier, and this was true irrespective of the tissue sampling method used (Fig. 5A-C). Sampling early in the lesion's evolution (e.g. near to the start of the simulations) had poor cor-

relation 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. 5B). 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 (analogous to endoscopic brushings or washings), the Shannon and Gini-Simpson indices showed good correlations with the time of clinically detectable cancer. These diversity measures were more correlated for random samples than for circular biopsies, despite each sample constituting similar numbers of cells (10% and 12% of the lesion, respectively). This result may reflect the fact that a biopsy 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 than a biopsy from a single time point. To do this, we evaluated the average of the values of each biomarker at the times t_1 and t_2, and the correlation between this average and the time of clinically detectable cancer. We then compared this correlation with that of the biomarker value at time t_2 alone. These results are shown in Fig. 6, 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 *average* of the individual biomarker values at each time points and the correlation coefficient for the biomarker value at the second time point alone.

Including information from an early biopsy in this manner provided slight additional prognostic value over-and-above the information available in the later biopsy (Fig. 6; approximately a 0.1 increase in the correlation was observed). In contrast, When information was combined by taking the difference in biomarker values between the biopsies collected at two different time points, the value from the later biopsy was generally more prognostic, and importantly was more prognostic than a measure which combined information both the early and late biopsies. Interestingly, the prognostic value of the Shannon and Gini-Simpson indices was reduced when considering the difference in biomarker values between two time points (Fig. S4). When we instead compared the maximal value of the biomarker across the two time points to its value at the later time point, we found similar results to the average case, but with smaller increases in correlation at later times; this was particularly the case for the Shannon, Gini-Simpson, IPP and INP indices (Fig. S5).

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 [42] and so cause an incorrect prognosis assignment. To address this issue, we studied how the prognostic value of each putative biomarker was improved by taking additional

biopsies at the same time ($T_b = 50$). For simplicity, after each virtual biopsy the sampled tissue was perfectly replaced in order to avoid the complexities associated with modelling local wound healing and tissue recovery. Further, while we did not strictly preclude biopsies from overlapping, the degree of overlap between biopsies is typically minimal because of the relatively small numbers of biopsies and small sizes of biopsy considered.

Assaying from more biopsies generally improved prognostic value, but with diminishing returns for each additional biopsy (Fig. 7). For all but one of the putative biomarkers, the maximum prognostic value was achieved by taking the average biomarker value across all biopsies, whereas measures of the spread of values (the variance or range) were generally poor prognosticators. Interestingly, the maximum prognostic value for the proportion of cells with at least two advantageous mutations was achieved by taking the *minimum* value across all biopsies; this could be because the 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 statistical association of each biomarker. We observed the same qualitative behaviour, such as diversity measures outperforming the proliferative fraction in degree of correlation, irrespective of the choice of parameter values or update rule used (see Supplementary Figs S1–S3, Supplementary Tables S1–S17 and Supplementary Text S1 for details).

To briefly summarize these results: (i) lower mutation rates decreased the correlation of each marker with the time of clinically detectable cancer, because of the increased stochasticity in the model introduced by a lower mutation rate; (ii) smaller biopsies were in general less prognostic; (iii) the number of mutations required for cancer did not qualitatively change the predictions of the model; (iv) the fitness advantage and disadvantage caused by new mutations,

and the relative likelihood of each of the various mutation types, did not qualitatively alter the prognostic value of the biomarkers, although diversity measures were most prognostic for the case where there were many strongly advantageous mutations; and (v) the closer the value of the threshold N_p was to the number of mutations required for cancer, the more correlated the proportion of cells with $n_p \geq N_p$ became with the time of clinically detectable cancer.

We also analyzed the sensitivity of model results to variations in the update rules of the system. We tested the biomarkers in a birth-driven system as opposed to a death-driven system, and found that again, that the diversity and IPP measures remained the biomarkers most significantly associated with the time of clinically detectable cancer. Further, we investigated the effect of decoupling mutations and cell division. While these changes to the model did alter the specific predictive values of the various indices (summarized in Table S14), the general pattern of statistical association was not altered. Moreover, even in this scenario, the IPP performed well.

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 predictors of the time of clinically detectable cancer, whereas measures of clonal diversity were highly correlated with the time of clinically detectable cancer and were robust to the choice of tissue sampling scheme. 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 than 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.

Averaging biomarker scores from two different time points did improve the predictive value of our putative biomarkers; however, the difference in each putative biomarker's values between time points was less predictive than its value at the later time points. This result was somewhat counter to our initial intuition that taking longitudinal biopsies would accurately track the 'evolutionary trajectory' of the lesion and hence dramatically improve prognostication. This result illustrates how our *in silico* approach can challenge intuition 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. It thus combines both genetic (or phenotypic) information with spatial (cell position) and dynamic (proliferation) information. This integration of multiple different sources of information may account for the prognostic value of the biomarker in our model. Empirical measurement of the IPP would be feasible if, for example, the number of driver mutations accumulated by a cell could be quantified concomitantly with a proliferative marker. 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 simple relationships between genotype, phenotype and fitness, and also neglected to model cell-cell interactions. 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 de-

scribing these kinds of important biological complexities within a model is a necessary next 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, although we have shown that our results are somewhat robust to alterations of a number of the key parameters in our model. Incorporating additional biological realism would also facilitate the *in silico* testing of the prognostic value of a full range of specific biological features; for example, the expression of a protein that fulfils 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 models 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

The authors wish to thank the anonymous reviewers for their insightful comments and suggested improvements to the manuscript.

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;17:3512–9.
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:S4–10.
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;365:1375–83.

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;24:3726–34.

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;304:1497–500.

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;64:3414–27.

10. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer. 2005;5:845–56.

11. Lari SA, Kuerer HM. Biological markers in DCIS and risk of breast recurrence: a systematic review. J Cancer. 2011;2:232–61.

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;12:323–34.

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;103:5923–8.

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;38:468–73.

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;120:636.

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;4:127ps10.

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;6:514–27.

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;77:42–8.

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;107:18545–50.

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;127:905–15.

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;14:371–80.

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;355:170–84.

30. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys. 1976;22:403–34.

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;14:2584–9.

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;12:159–68.

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. Scholzen T, Gerdes J. The Ki-67 protein: from the known and the unknown. J Cell Physiol. 2000;182:311–22.

35. Shannon CE. Communication theory of secrecy systems. Bell Syst Tech J. 1949;28:656–715.

36. Simpson EH. Measurement of diversity. Nature. 1949;163:688.

37. Jost L. Entropy and diversity. Oikos. 2006;113:363–75.

38. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37:17–23.

39. Geary RC. The contiguity ratio and statistical mapping. The Incorporated Statistician. 1954;5:115–46.

40. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004;159:882–90.

41. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61:92–105.

42. 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;366:883–92.

Tables

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	Min. no. 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	$\{0.01, 0.05, 0.1\}$
N_p	Min. no. advantageous mutations for positive stain	$\{2, 3, 5, 7, 9\}$
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	$\{50, 80\}$

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' denotes the change in the value of each putative biomarker that increases the associated hazard ratio by the reported factor.

		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)	$< 10^{-4}$
	Gini-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)	$< 10^{-4}$
	Gini-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	$\mathbf{< 10^{-6}}$	(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)	$< 10^{-4}$
	Gini-Simpson index	0.01	**3.5**	(2.8, 4.4)	$< 10^{-4}$

Figure legends

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 averaged from 200 simulations with parameter values $N_m = 10$, $s_p = s_d = 0.2$ for five such genotypes (for figure clarity).

Figure 2

Prognostic value of random tissue sampling. A 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 Gini-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.

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 the mitotic proportion (A), Shannon index (B), Gini-Simpson index (C), Moran's *I* (D), Geary's *C* (E), IPP (F) and INP (G). In each case, 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.

Figure 4

Areas under ROC curves for putative biomarkers. The prognostic value of sampling a circular biopsy at time $T_b = 75$ was assessed by considering the area under the curve (AUC) of receiver-operator characteristic (ROC) curves as a function of censoring time. This analysis confirmed the time-invariant predictive value of the IPP (red line) and clonal diversity measures (blue and green lines), and lack of predictive value derived from the mitotic proportion (black line) and proportion of cells bearing at least one abnormality (brown line). The worse-than-random performance of the proliferation and Geary's *C* measures at short censoring times is likely to be attributable to the stochasticity inherent in cancer development within the model: early clonal expansions do not necessarily signify later cancer risk. Results from 1000 simulations for each sampling scheme, with parameter values $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$,

$N_b = 20$ and $N_s = 10^3$. For comparison, the black dotted line denotes an AUC = 0.5 (which would would be achieved by a random predictor).

Figure 5

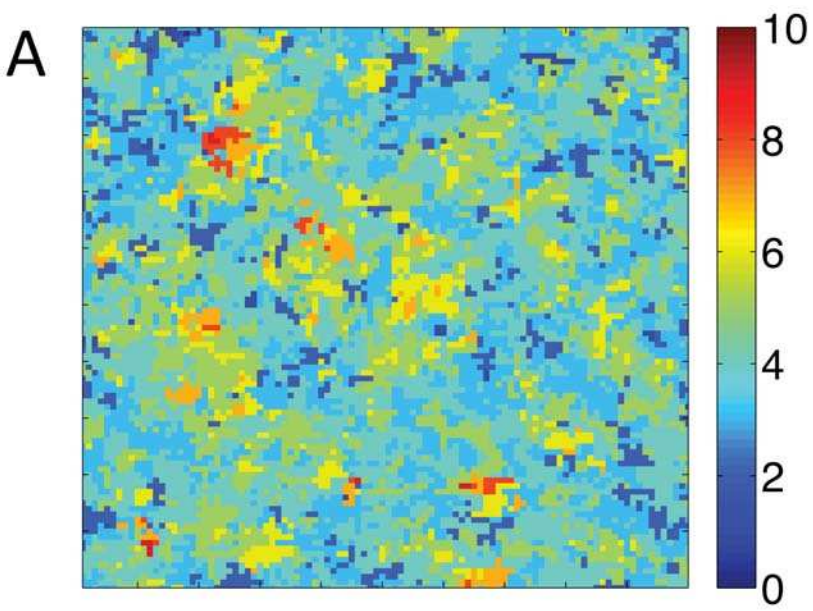
Prognostic value of early versus late biopsies. For a range of sampling times T_b, the virtual tissue was biopsied and the correlation between putative biomarker values and the time of clinically detectable cancer was computed. Results are shown based on sampling the whole lesion (A), a circular biopsy (B) and random tissue sampling (C). For each sampling scheme, 1000 simulations were run with $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$, $N_b = 20$ and $N_s = 10^3$.

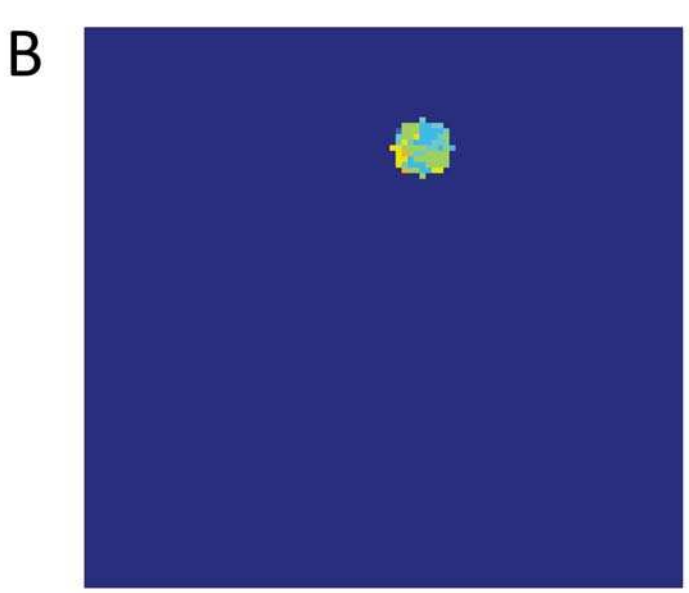
Figure 6

Serial biopsies provide slightly increased additional prognostic information. Heat maps depicting the relative value of taking serial biopsies at different time points for the proportion of cells with at least two positive mutations (A), mitotic proportion (B), Shannon index (C), Gini-Simpson index (D), Moran's *I* (E), Geary's *C* (F), IPP (G) and INP (H). Positive values (warm colours) indicate that prognostic value was improved by taking the average of biomarker value from both time-points; negative values (cool colours) indicate that more information was available at the second time point alone than from the averaged time points. Results are shown from 1000 simulations for each pair of time points, with $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$ and $N_b = 20$.

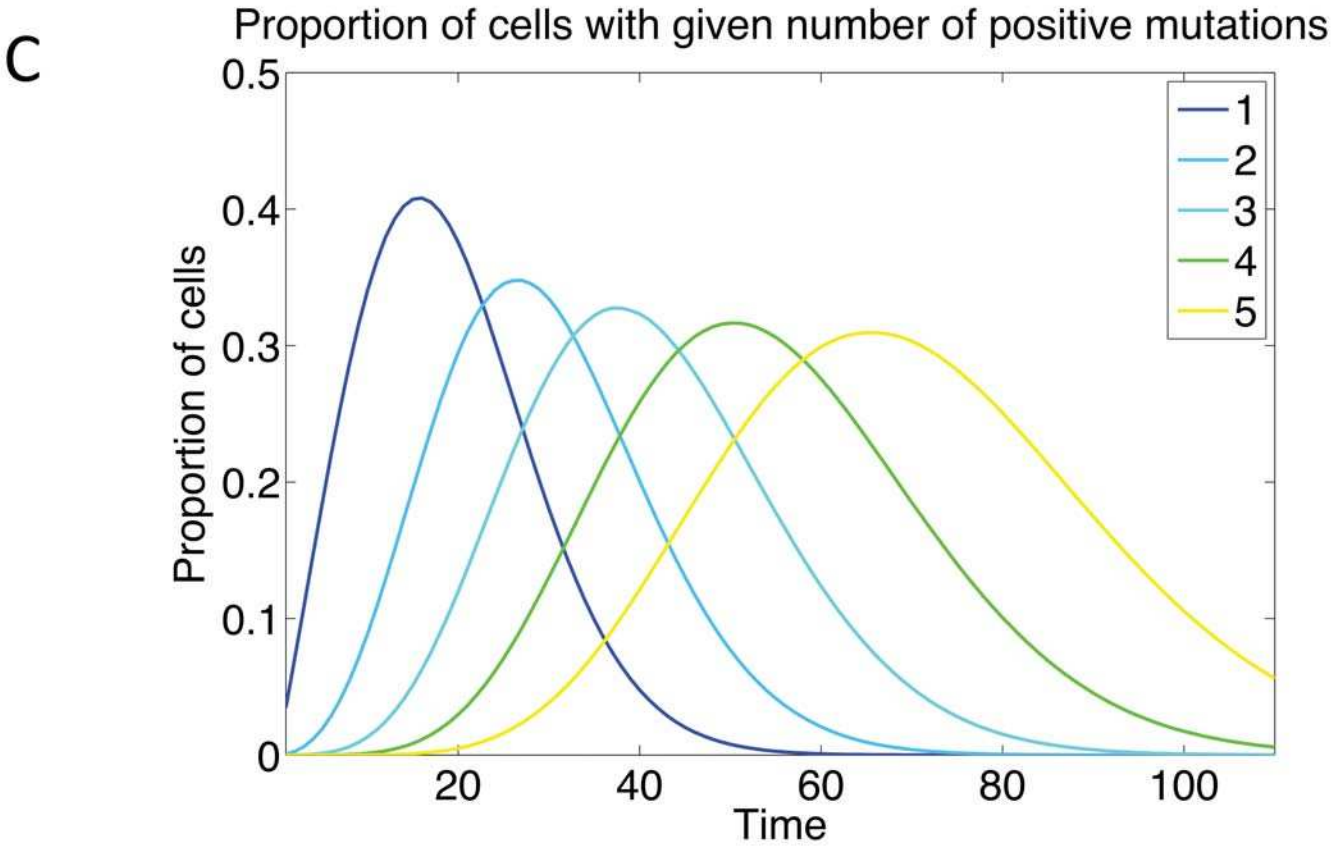
Figure 7

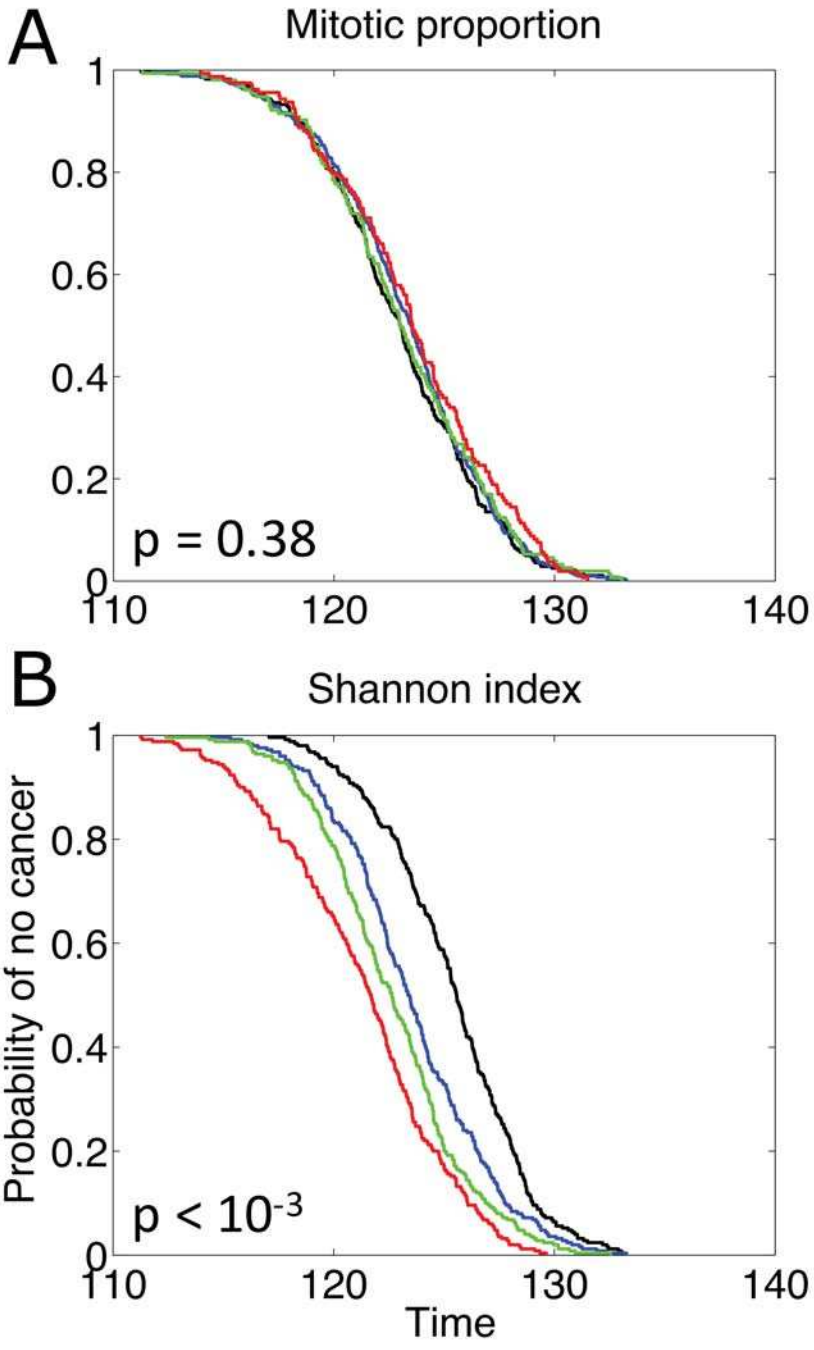
Additional biopsies at the same time point improves prognostication with diminishing returns. Graphs show the relationship between the correlation coefficient (between each biomarker value and time of clinically detectable cancer) and the number of biopsies collected at time $T_b = 50$, for the proportion of cells with at least two positive mutations (A), mitotic proportion (B), Shannon index (C), Gini-Simpson index (D), Moran's *I* (E), Geary's *C* (F), IPP (G) and INP (H). 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). Results are shown from 1000 simulations for each pair of time points, with $N_m = 10$, $s_p = s_d = 0.2$, $\mu = 0.1$ and $N_b = 20$.

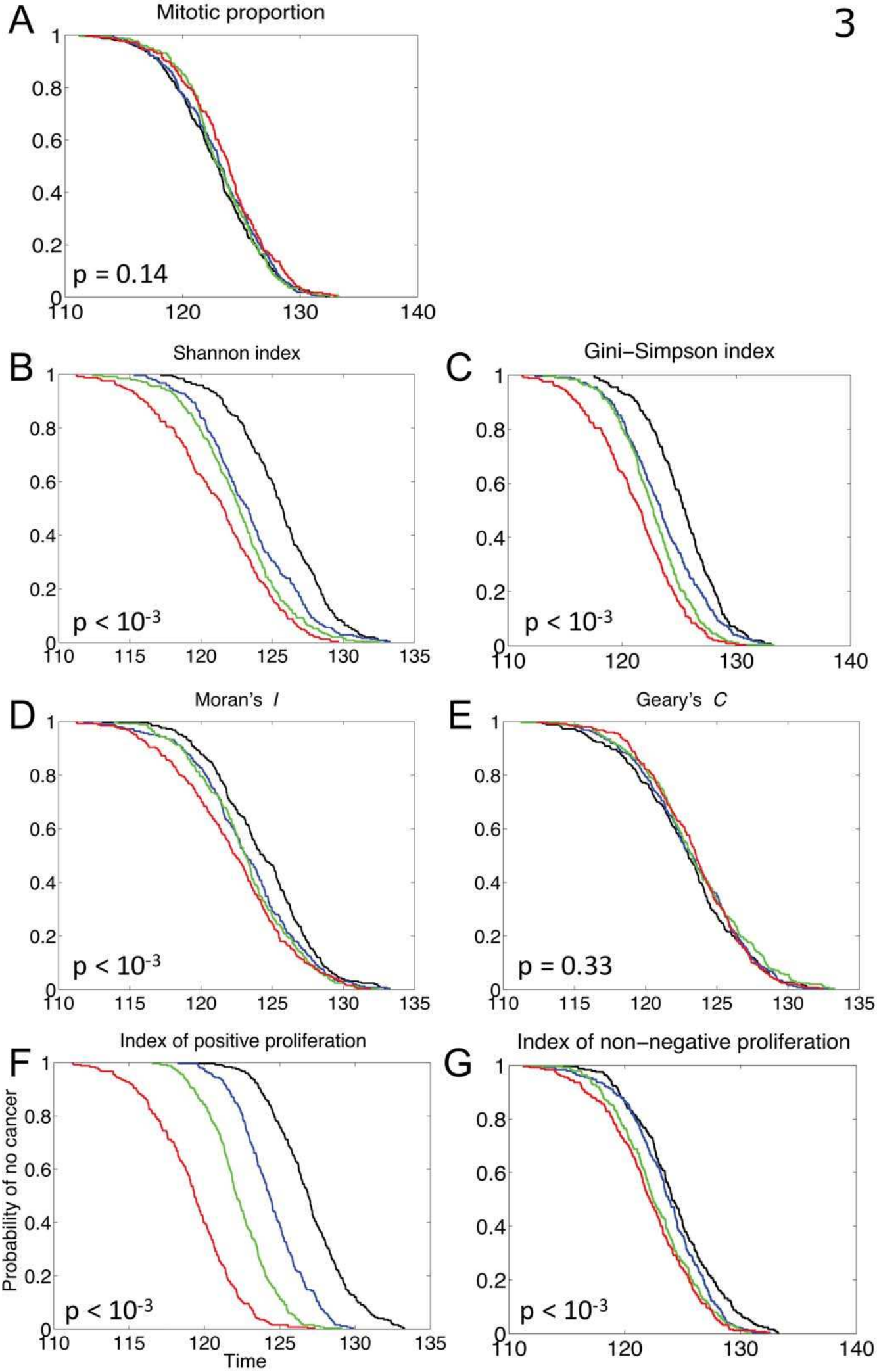

A
10
8
6
4
2
0

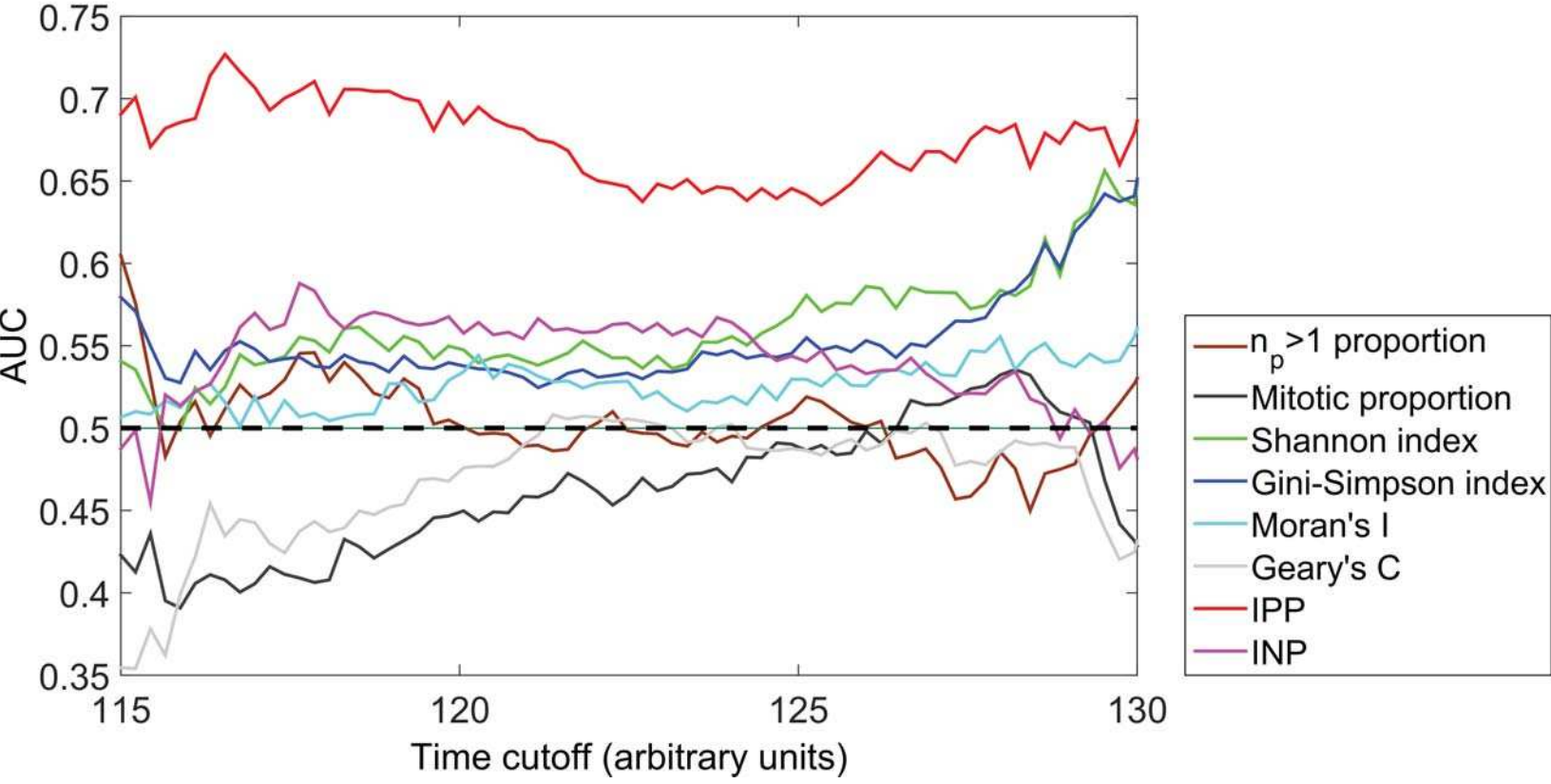

B

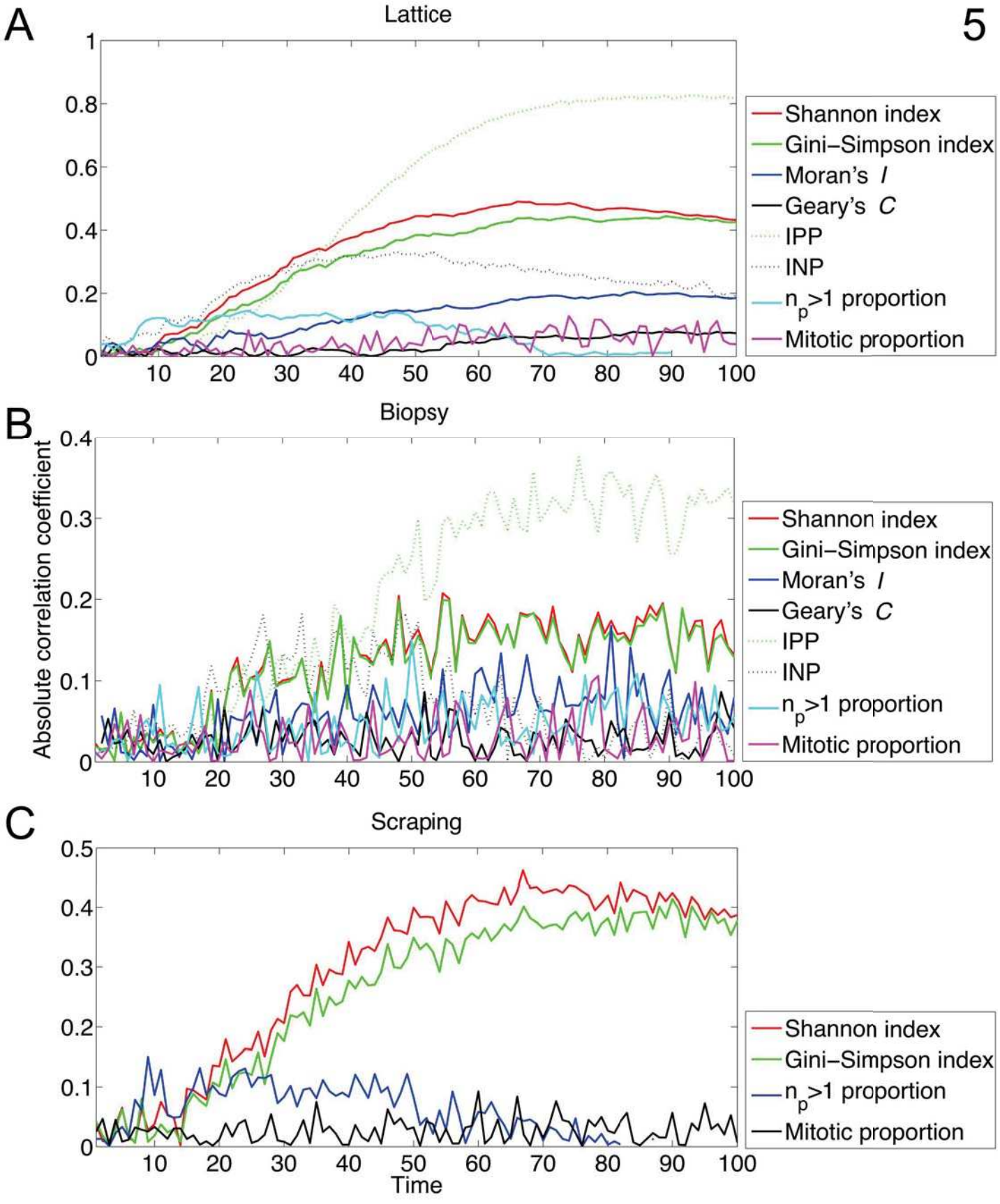
1

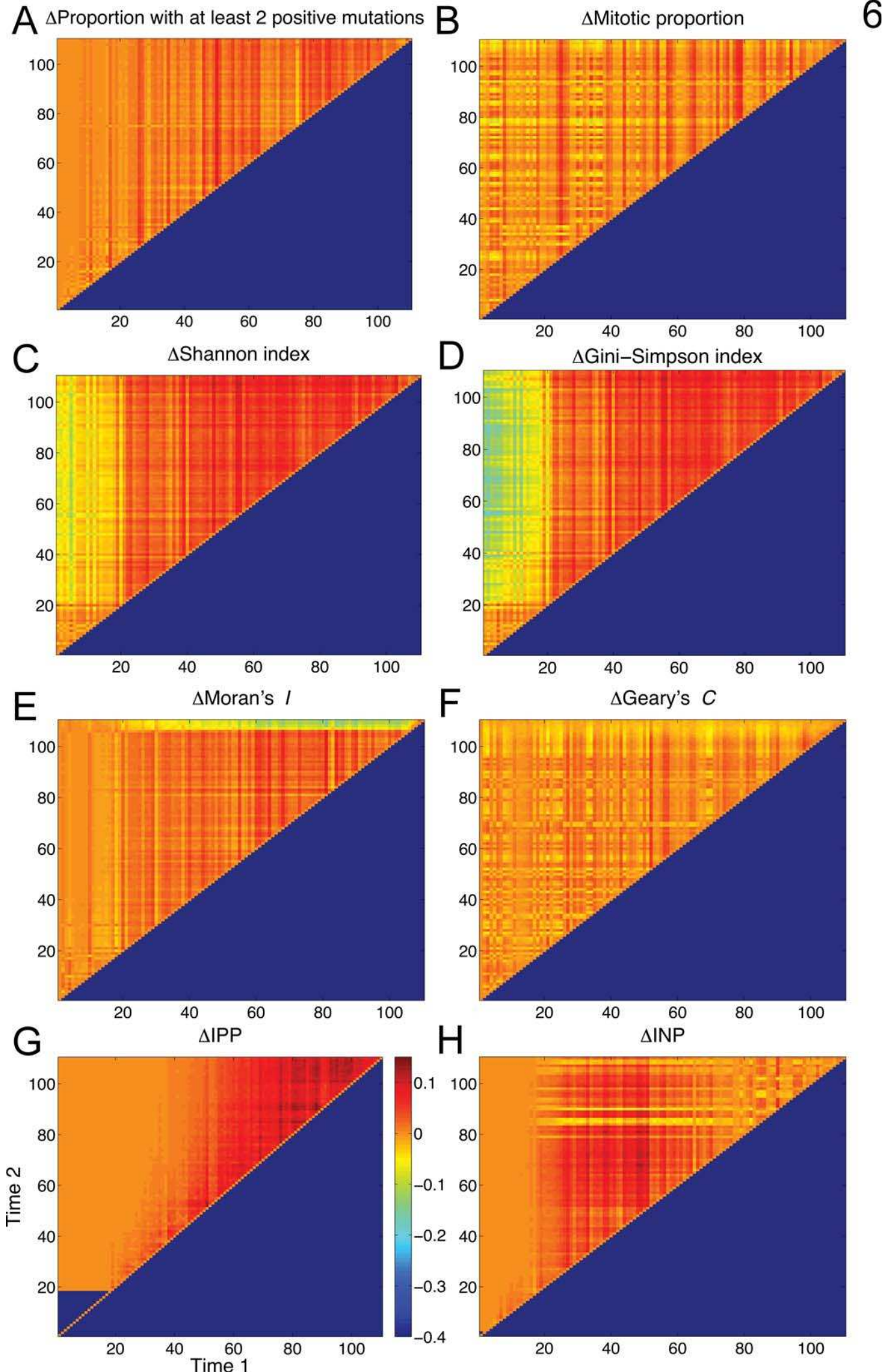

C
Proportion of cells with given number of positive mutations
Proportion of cells
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
Time
1
2
3
4
5

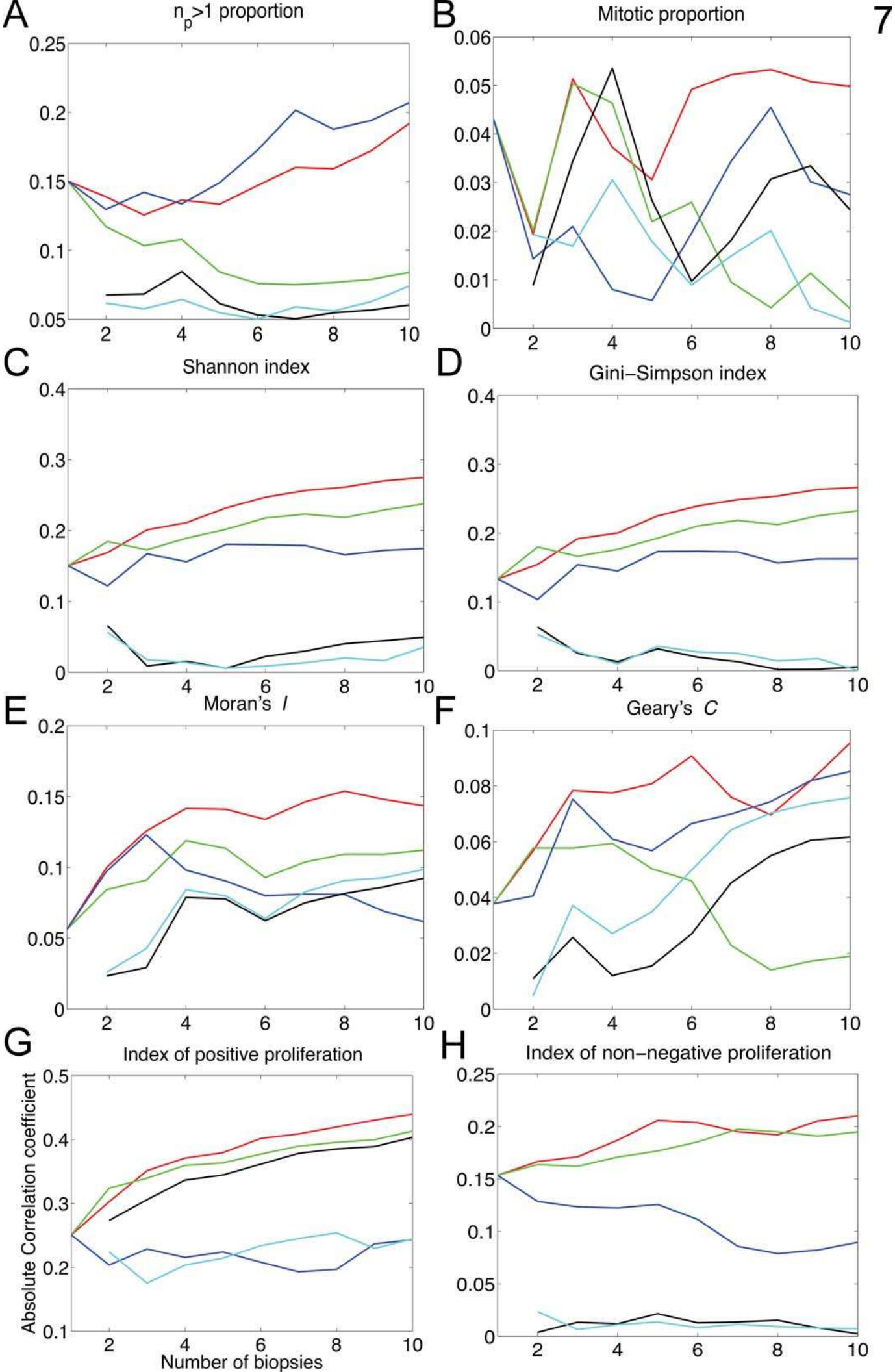

A
Mitotic proportion
1
0.8
0.6
0.4
0.2
0
p = 0.38
110
120
130
140
B
Shannon index
Probability of no cancer
p < 10-3
Time
C
Gini–Simpson index
p < 10-3


A
Mitotic proportion
p = 0.14
B
Shannon index
p < 10-3
C
Gini–Simpson index
p < 10-3
D
Moran's I
p < 10-3
E
Geary's C
p = 0.33
F
Index of positive proliferation
p < 10-3
G
Index of non–negative proliferation
p < 10-3
Probability of no cancer
Time


0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
AUC
115
120
125
130
Time cutoff (arbitrary units)
n_p>1 proportion
Mitotic proportion
Shannon index
Gini-Simpson index
Moran's I
Geary's C
IPP
INP


A
Lattice
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
100
Shannon index
Gini–Simpson index
Moran's I
Geary's C
IPP
INP
n_p>1 proportion
Mitotic proportion
B
Biopsy
Absolute correlation coefficient
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
Shannon index
Gini–Simpson index
Moran's I
Geary's C
IPP
INP
n_p>1 proportion
Mitotic proportion
C
Scraping
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
Time
Shannon index
Gini–Simpson index
n_p>1 proportion
Mitotic proportion


A
ΔProportion with at least 2 positive mutations
B
ΔMitotic proportion
C
ΔShannon index
D
ΔGini–Simpson index
E
ΔMoran's I
F
ΔGeary's C
G
ΔIPP
H
ΔINP
Time 2
Time 1
20
40
60
80
100
0.1
0
−0.1
−0.2
−0.3
−0.4


A
n_p>1 proportion
B
Mitotic proportion
C
Shannon index
D
Gini–Simpson index
E
Moran's *I*
F
Geary's *C*
G
Index of positive proliferation
H
Index of non–negative proliferation
Absolute Correlation coefficient
Number of biopsies