Multifocality and recurrence risk: a quantitative model of field cancerization

Jasmine Foo[1*], Kevin Leder [2*], and Marc D. Ryser[3†]
1. School of Mathematics, and 2. Industrial and Systems Engineering,
University of Minnesota, Minneapolis, MN
3. Department of Mathematics, Duke University, Durham, NC

February 13, 2014

Abstract

Primary tumors often emerge within genetically altered fields of premalignant cells that appear histologically normal but have a high chance of progression to malignancy. Clinical observations have suggested that these premalignant fields pose high risks for emergence of recurrent tumors if left behind after surgical removal of the primary tumor. In this work, we develop a spatio-temporal stochastic model of epithelial carcinogenesis, combining cellular dynamics with a general framework for multi-stage genetic progression to cancer. Using the model, we investigate how various properties of the premalignant fields depend on microscopic cellular properties of the tissue. In particular, we provide analytic results for the size-distribution of the histologically undetectable premalignant fields at the time of diagnosis, and investigate how the extent and geometry of these fields depend upon key groups of parameters associated with the tissue and genetic pathways. We also derive analytical results for the relative risks of local vs distant secondary tumors for different parameter regimes, a critical aspect for the optimal choice of post-operative therapy in carcinoma patients. This study contributes to a growing literature seeking to obtain a quantitative understanding of the spatial dynamics in cancer initiation.

1 Introduction

The term 'field cancerization' refers to the clinical observation that certain regions of epithelial tissue have an increased risk for the development of multiple synchronous or metachronous primary tumors. This term originated in 1953 from repeated observations

*Partially supported by NSF grant DMS-1224362
†Partially supported by NIH grant R01-GM096190-02

by Slaughter and colleagues of multiple primary oral squamous cell cancers and local recurrences within a single region of tissue [1]. The phenomenon, also known as the 'cancer field effect' has been documented in many organ systems including head and neck (oral cavity, oropharynx, and larynx), lung, vulva, esophagus, cervix, breast, skin, colon, and bladder [2]. Although the exact underlying mechanisms of the field effect in cancer are not fully understood, recent molecular genetic studies suggest a carcinogenesis model in which clonal expansion of genetically altered cells (possibly with growth advantages) drives the formation of a premalignant field [2, 3]. This premalignant field, which may develop in the form of one or more expanding patches, forms fertile ground for subsequent genetic transformation events, leading to intermediate cancer fields and eventually clonally diverging neoplastic growths. The presence of such premalignant fields poses a significant risk for cancer recurrence and progression even after removal of primary tumors. Importantly, these fields with genetically altered cells often appear histologically normal and are difficult to detect; thus, mathematical models to predict the extent and evolution of these fields may be useful in guiding treatment and prognosis prediction.

In this work we utilize a stochastic evolutionary framework to model the cancer field effect. Our model combines spatial cellular reproduction and death dynamics in an epithelial tissue with a general framework for multi-stage genetic progression to cancer. Using this model, we investigate how microscopic cellular properties of the tissue (e.g. tissue renewal rate, mutation rate, selection advantages conferred by genetic events leading to cancer, etc) impact the process of field cancerization in a tissue. We develop methods to characterize the waiting time until emergence of second field tumors and the recurrence risk after tumor resection. In addition we study the clonal relatedness of recurrent tumors to primary tumors by assessing whether local field recurrences (second field tumors) are more likely than distant field recurrences (second primary tumors). The key results of our study are summarized as follows. (i) We provide analytic results for the size-distribution of the histologically undetectable pre-cancerous fields at the time of diagnosis. (ii) We investigate how the extent and geometry of these fields depend upon a key meta-parameter of the system, Γ, which is defined through a specific relationship between kinetic parameters of the tissue and genetic pathways. (iii) We derive analytical results for the relative risks of local vs distant secondary tumors for different parameter regimes. These types of predictions are important in clinical practice. For example, they help determining the optimal size of excision margins at the time of surgery, and the appropriate choice of post-operative therapy (which may depend on the type of recurrence expected).

The methodology developed in this work is generally applicable to early carcinogenesis in epithelial cancers, and contributes to a growing literature on the evolutionary dynamics of cancer initiation, see e.g. [4–13]. Since our work is concerned with analyzing spatial premalignant field geometries during the genetic progression to cancer, here we briefly describe some existing mathematical models of the stochastic evolutionary process of cancer initiation from spatially structured tissue, e.g. [14–19]. In 1977 Williams and Bjerknes proposed a spatial Moran model of clonal expansion in epithelial tissue [16] in which cells

divide according to fitness and replace a neighboring cell at random on the rectangular lattice. This model is closely related to the biased voter model from particle systems theory [20], and in [21, 22] the growth properties and asymptotic shape of the process were established. However, this model did not incorporate the possibility of mutations occurring to produce new types in the population. In [14] Komarova proposed a 1D model incorporating mutations with fitness advantages, where cells were allowed to divide in response to the death of a neighboring cell in contrast to the models mentioned previously. It was shown that the probability of mutant fixation and time to obtain two-hit mutants differ from the well-mixed setting. Later, in [17,18] this model was extended to incorporate motility, and the relationships between migration, mutation, selection and invasion in a spatial stochastic evolutionary model were explored. In [19] the voter model considered in [16] was generalized to incorporate neutral mutations, and the waiting time to produce two-hit mutants was studied in a general dimension setting. Martens and colleagues considered a similar model of mutation accumulation on a discrete time hexagonal lattice model, and studied the speed of population adaptation [23, 24]. In a recent work Antal and colleagues consider a stochastic spatial model of cancer progression where cells acquire successive fitness advantages along the edge of the tumor. In the context of this model they study the shape of the evolving tumor front as well as the number of mutations acquired in the tumor [25]. In a recent work, we studied the accumulation and spread rates of advantageous mutant clones in a spatially structured population of general dimension [26]. Finally, we note that there have also been some studies mathematically modeling the growth of pre-cancerous cells via growth factors during early carcinogenesis utilizing reaction-diffusion systems, e.g. [27].

Most of the evolutionary models proposed in the field utilize similar descriptions of the fundamental processes of birth, selection, mutation and death in a spatially structured population (modulo the occasional minor differences in lattice structure and the structure of reproduction update rules). However, the studies described above have been aimed at studying the rates of invasion, adaptation, and mutation accumulation in these populations. In contrast, in this study we obtain analytical results for the spatial and temporal dynamics of premalignant fields during carcinogenesis. We consider a generalized spatial Moran process in which cells can acquire successive random mutations which confer selective advantages, reproduction occurs at rates proportional to cellular fitness, and reproduction results in neighbor replacement at random. We analyze this fundamental evolutionary model to quantify how field cancerization dynamics and recurrence risks depend on the kinetic parameters of the tissue and genetic progression pathway to cancer. To the best of our knowledge, this is the first evolutionary modeling effort aimed at mathematically predicting the cancer field effect and its consequences.

The article is organized as follows: in section 2 we introduce the stochastic mathematical model and describe basic properties regarding the survival and growth rate of mutant clones. Using previously derived results on the spread of mutant clones, we introduce a mesoscopic approximation to the model. In section 3 we analyze the model to investigate

the characteristics and extent of local and distant premalignant fields at the time of initiation. In particular, we determine how the spatial geometry of the field (e.g. number and size of lesions) depends on cellular and tissue properties such as mutation rate, tissue renewal rate and mutational fitness advantages. In section 4 we analyze the model to understand the risk of recurrence due to local or distant field malignancies, as a function of time and cellular parameters.

Throughout the paper we will use the following notation for the asymptotic behavior of positive functions,

$$\begin{aligned} f(t) \sim g(t) &\quad \text{if } f(t)/g(t) \to 1 \text{ as } t \to \infty, \\ f(t) \ll g(t) &\quad \text{if } f(t)/g(t) \to 0 \text{ as } t \to \infty, \\ f(t) \gg g(t) &\quad \text{if } f(t)/g(t) \to \infty \text{ as } t \to \infty. \end{aligned}$$

Finally, we use the notation $X =_d F$ to denote that the random variable X has distribution F.

2 Mathematical framework and basic properties

Cancer initiation is associated with the accumulation of multiple successive genetic or epigenetic alterations to a cell [28]. A subset of these genetic events may give rise to a fitness advantage (i.e. an increase in reproductive rate of the cell or avoidance of apoptotic signals), and subsequently lead to a clonal expansion within the tissue. These expanding mutant cell populations form the background for further independent genetic events which eventually lead to carcinogenesis. As a result of this spatial evolutionary process, by the time of cancer initiation or diagnosis the tissue field surrounding a tumor can be composed of genetically distinct premalignant lesions of various sizes and stages.

2.1 Cell-based model

To study the dynamics of this process, we consider a stochastic model which describes the accumulation and spread of a clone of cells with genetic alterations throughout a spatially structured tissue (e.g. stratified epithelium). Thus, we consider the model on a regular lattice $\mathbb{Z}^d \cap [-L/2, L/2]^d$, where $L > 0$ and d is the number of spatial dimensions of the tissue. Each location in the lattice is occupied by a single cell, and each cell reproduces at a rate according to its fitness with exponential waiting times. Whenever a cell reproduces, its offspring replaces one of its $2d$ lattice neighbors at random, see Figure 1A. The type of each cell corresponds to its fitness, which is related to the number of genetic hits a cell has accumulated in a multi-step genetic model of cancer initiation. For example, type-0 cells have fitness normalized to 1 and are labeled as wild-type or normal (with no mutations). Initially our entire lattice is occupied by type 0 cells. Type-0 cells acquire the

first mutation at rate u_1 to become type-1 cells. The type-1 cell will have a relative fitness advantage to type-0 cells, given by $1 + s_1$, for some constant $s_1 \geq 0$. In general, type-i cells have a fitness advantage of $1 + s_i$ relative to type-$(i-1)$ cells, and they acquire the $(i+1)-$th mutation in the sequence at rate u_{i+1} to become type-$(i+1)$ cells. The process is stopped when a cell develops k mutations; we call this the time of cancer initiation. The number of mutation k as well as the parameters u_i, s_i for $i = 1, \ldots, k$ depend on the specific cancer type. Although many (epi)genetic events are selectively disadvantageous (i.e. they confer a selective disadvantage $s_i < 0$), the progeny of deleterious mutants die out quickly so here we restrict our attention to the case $s_i \geq 0$. Note that this process can be thought of as a spatial version of the Moran process, a spatially well-mixed population model that is commonly used to describe carcinogenesis (e.g. see [8–12]). In addition, the spatial reproduction and death dynamics of this model (without mutation) correspond to the biased voter process which has been well-studied in physics and probability literature. In fact, a similar voter model approach was previously used to model cellular dynamics in epithelial tissue and found to correlate well with experimental predictions of clone size distribution in the mouse epithelium [29].

The total number of cells in the fixed-size population is $N \equiv L^d$; in most cancer initiation settings this number is quite large (at least 10^6), while mutation rates are quite small (orders of magnitude smaller than 1). Therefore we will, unless stated otherwise, restrict our analysis to regimes where $L \gg 1$ and $u_i \ll 1$. In Section 2.3, we will briefly discuss the specific conditions that we impose on the relationship between these parameters. For mathematical simplicity, the lattice is equipped with periodic boundary conditions; however in most relevant biological situations the domain size (i.e. cell number) is sufficiently large so that boundary effects are negligible.

Note on dimension of the model. We analyze the general model in space dimensions $d = 1, 2, 3$. While all epithelial tissues have an intrinsically three dimensional architecture, in some situations considering $d = 1, 2$ may be a good approximation. For example, cancer initiation in mammary ducts of the breast, renal tubules of the kidney, and bronchi tubes of the lung could be viewed as approximately one-dimensional processes, due to the aspect ratio of tube radius versus length. On the other hand, cancer initiation in the squamous epithelium of the cervix, the bladder or the oral cavity can be viewed as two-dimensional process, since initiation occurs in the basal layer of the epithelium which is only 1-2 cells thick (see e.g. Figure 2). The validity of such approximations poses an interesting problem in itself, but will not be addressed in this work.

2.2 Survival and growth of a single mutant clone

We first establish some basic behaviors of mutant cells and their clonal progeny within a tissue. Of particular interest are: (i) the survival probability of a mutant clone, and (ii) the rate of spatial expansion of the mutant clone through the tissue. In particular, how are these characteristics influenced by tissue parameters and the cellular fitness advantage

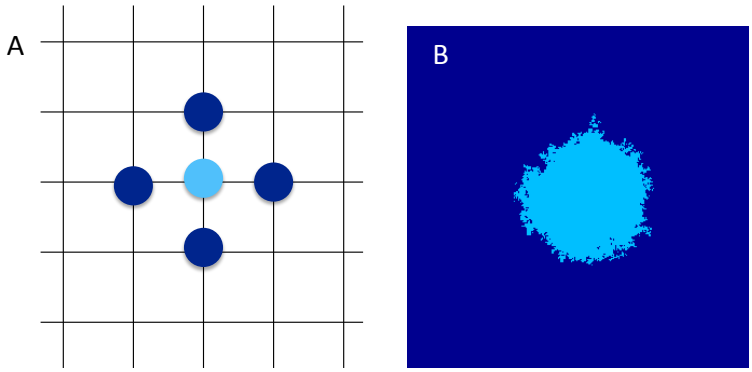

Figure 1: **Lattice dynamics.** (A) Schematic of spatial Moran model in $d = 2$: each cell divides at rate according to its fitness and replaces one of its $2d$ neighbors: if the light blue cell divides, its offspring replaces one of the dark blue neighbors, chosen uniformly at random. Every lattice site is occupied at all times (not shown). (B) Simulation example of the model: growth of an advantageous clone (light blue) starting from one cell with fitness advantage $s = 0.2$ over the surrounding field (dark blue).

conferred by a mutation? We have addressed some of these questions in a previous work [26] and restate the results here to make the paper self-contained. In addition, we perform new simulations in this work to fill in gaps where theoretical results are currently not available.

Consider the probability that a mutant cell survives to form a viable clone (i.e. does not die out due to demographic stochasticity). Let type-1 cells have fitness $1 + s$ and type-0 cells have fitness 1, and let $\phi_t(x)$ denote the type of cell at site x in the lattice at time t. Define

$$\xi_t \equiv \{x \in \mathbb{Z}^d \cap [-L/2, L/2]^d : \phi_t(x) = 1\}.$$

In other words, ξ_t is the set of all type-1 cell locations at time t. We initiate the model with a single type-1 cell at the origin surrounded by type-0 cells in all other locations:

$$\phi_0(x) = \begin{cases} 1, & x = 0 \\ 0, & \text{otherwise,} \end{cases}$$

and assume no further mutations are possible ($u_i = 0$). This simplified model is known as the Williams-Bjerknes model [16], and if $L = \infty$ then it corresponds to the biased voter model, see e.g. [30]. Let $|\xi_t|$ denote the number of type-1 cells in the model at time t. Then we can define the extinction time of the process $T_0 \equiv \inf\{t > 0 : |\xi_t| = 0\}$. The probability of survival of a single mutant clone with selective advantage s over the surrounding cells is then the probability of the event $\{T_0 = \infty\}$. By looking at the the process $|\xi_t|$ only at its

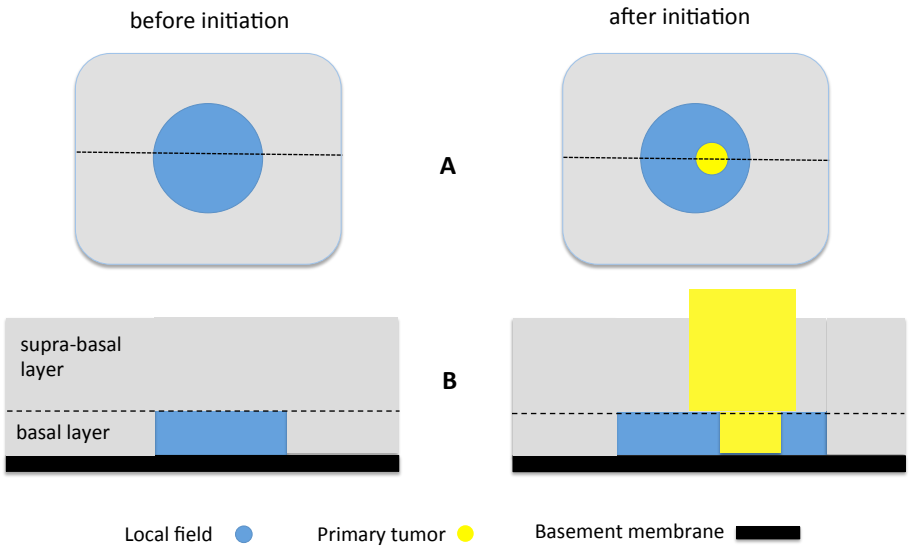

Figure 2: **Geometry of squamous epithelium.** **A** Basal layer (vertical perspective) before initiation with local field (left), and after initiation where the tumor is growing within the local field (right). **B** Sideways view of the fields before and after initiation, along the dashed lines in panel A. The proliferative cells inhabiting the two-dimensional lattice in the model reside in the basal layer of the epithelium.

jump times, we note that the embedded process is a discrete time random walk that moves one up with probability $s/(1+s)$ or one down with probability $1/(1+s)$. This can be seen by observing that the process only changes at boundaries between type-0 and type-1 cells, and the only possible resulting events are that the type-0 gets replaced by a type-1 (resulting in a jump up in $|\xi_t|$) or the type-0 gets replaced by a type-1 (resulting in a jump down in $|\xi_t|$). Analysis of the overall survival probability of this random walk can then be calculated using elementary results for random walks, see Example 1.43 in [31],

$$P(T_0 = \infty) = \frac{s}{1+s} \approx s,$$

where the approximation is valid for $s \ll 1$. Thus, the probability that a mutant clone with fitness advantage s survives is $\frac{s}{1+s}$, and is independent of the dimension of the tissue.

To understand how the expansion rate of a mutant clone depends on the selection strength s of the mutant, we first recall a result by Bramson and Griffeath [21, 22], which establishes an asymptotic shape for the type-1 clone. More precisely, Bramson-Griffith shape theorem says that conditional on the clone never going extinct, the clone has a convex, symmetric shape whose radius expands linearly. In a previous work, we studied how this linear rate of expansion depends on the selection strength s in the setting of weak selection, see Theorem 1 of [26]. We found that if we denote by e_1 the first unit vector in $\mathbb{R}^d$ and define the growth rate $c_d(s)$ such that

$$D \cap \{ze_1 : z \in \mathbb{R}\} = [-c_d(s), c_d(s)],$$

then as $s \to 0$,

$$c_d(s) \sim \begin{cases} s & d = 1 \\ \sqrt{4\pi s/\log(1/s)} & d = 2 \\ \sqrt{4\beta_3 s} & d = 3, \end{cases} \tag{1}$$

where β_3 is the probability that two simple random walks started at 0 and $e_1 = (1, 0, 0)$ never hit. In other words, the radius of the asymptotic shape D approximating the type-1 clone grows linearly with rate on the order of $c_d(s)$.

The previous results hold only in the regime of weak selection or small s. For larger values of the selective advantage s, simulations can be used to obtain $c_d(s)$ for $d = 2, 3$ (in $d = 1$ the process can be analyzed directly through simple random walk analysis and we obtain that $c_1(s) = s$). For example, Figure 3 shows that the s-dependence of the growth rate is approximately linear for $s > 0.5$; in this case simple regression yields the estimate $c_2(s) \approx 0.6s + 0.22$ $(s > 0.5)$. Thus, a combination of analysis and simulation gives us a complete picture of how spatial expansion rate of mutant clones in a tissue depend upon the selective advantage s for a wide range of selection strengths.

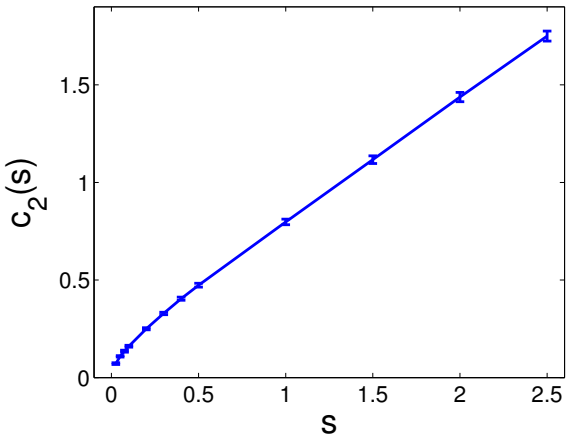

Figure 3: **Simulations of clonal expansion rate for large s.** Dependence of the growth rate c_2 on the fitness advantage s. Statistics performed on $M = 100$ samples for each s-value. The error bars represent 95% confidence intervals.

2.3 Approximating with a hybrid mesoscopic model

Our results regarding the survival and growth of a single mutant clone suggest a hybrid mesoscopic model simplification that enables our analysis of the field cancerization process. In particular, each successful mutant clone can be well-approximated as a growing d-dimensional ball with expansion rate $c_d(s)$ as calculated in the previous section. Before proceeding however, let us clarify the notion of clone 'survival' a.k.a. 'success' in the full model, where multiple mutations can arise and compete in the same finite domain. In particular, we consider a mutant clone with selective advantage s over the background to be successful if it reaches size $\gg 1/s$. This criterion guarantees a negligible chance of extinction in an infinite domain with no interference. In particular, if we start with a single type-1 cell with selective advantage s in a sea of type-0 cells, and if we define T_0 to be the extinction time of the type-1 progeny, one can use the embedded discrete time process and standard results on biased random walks [31] to show that if the progeny reaches size $k \gg 1/s$, then $P\left(T_0 = \infty \middle| |\xi_0| = k\right) \approx 1 - e^{-ks}$.

Consider the fate of an unsuccessful type-1 clone arising on a background of type-0 cells. The clone evolves as a supercritical ($s > 1$) biased voter model conditioned on extinction. In [26] we showed that unsuccessful type-1 mutations typically die out by a time of order

$$\ell(s) = \begin{cases} s^{-2} & d = 1, \\ s^{-1} \log(1/s) & d = 2, \\ s^{-1} & d = 3. \end{cases} \tag{2}$$

As seen in the previous section, the survival probability in the biased voter model (starting

with a single type 1 cell in a sea of type 0 cells) is $s/(1+s)$, but in the more complex spatial Moran model with the possibility of multiple interacting type 1 clones, it is not immediately clear that this survival probability is still given by $s/(1+s)$. However, it was shown in [26] that the above survival probability remains a good approximation as long as

$$(A0) \qquad (1/u_1) \gg \ell(s)^{(d+2)/2}. \tag{3}$$

If the total number of type-1 cells is always a negligible fraction of N and (A0) holds, then successful type-1 mutations arrive as a Poisson arrival process with approximate rate $Nu_1\frac{s}{s+1}$, where N is the total number of cells in the tissue. In particular, these conditions hold for biologically reasonable parameter sets, such as the ones used for the numerical examples in this article.

We are now ready to introduce a hybrid mesoscopic model approximation as follows: Type-1 mutations arrive in the healthy tissue as a Poisson arrival process with rate Nu_1, distributed uniformly at random in the spatial domain. Each mutation event has two potential outcomes:

- with probability $s/(1+s)$, the mutation is successful and we approximate the subsequent clonal expansion with a ball whose radius grows deterministically. The macroscopic growth rate is $c_d(s)$, which was derived from individual cellular growth kinetics as described in section 2.2. As a representative simulation in figure 1B suggests, the ball in standard L^2-norm in $\mathbb{R}^d$ will be utilized.

- with probability $1/(1+s)$, the mutation is unsuccessful, and the clone evolves according to the full stochastic (cellular-level) model dynamics conditioned on extinction.

Note that the remainder of the paper discusses properties of this mesoscopic model.

It will be useful to define γ_d as the volume of a ball of radius 1 in d dimensions,

$$\gamma_1 = 2, \qquad \gamma_2 = \pi, \qquad \gamma_3 = 4\pi/3.$$

Note that although the stochastic fluctuations of the shape of expanding clones are lost in this approximation, one gains generality since the mesoscopic model can approximate a whole class of microscopic models that admit a shape result.

2.4 Cancer initiation behavior

Although the methodology developed in this work can be generalized to the setting of k-mutation carcinogenesis models, we will consider for simplicity the classic two-mutation model of cancer initiation first introduced by Knudson [32]. Here, type-0 cells are wild-type with fitness 1, type-1 cells are premalignant with fitness $1+s_1$ relative to type-0 cells, and type-2 cells are initiated cancer cells with fitness $1+s_2$ relative to type-1 cells. The time of cancer initiation σ_2 is defined as the time at which the first successful type-2 cell arrives.

In [26], we studied the situation where $s_1 = s_2 = s > 0$ and found that the timing of cancer initiation is strongly governed by the limiting value of the following meta-parameter:

$$\Gamma \equiv (Nu_1 s)^{d+1}(c_d^d u_2 s)^{-1}.$$

Roughly speaking, $\Gamma^{1/(d+1)}$ represents the ratio of the rate of producing successful type-1 cells to the subsequent time it take to acquire the first successful type-2. We found that both the mechanisms and distribution of the cancer initiation time vary significantly depending on the regime of Γ:

- Regime 1 (R1): When $\Gamma < 1$, the first successful type-2 mutation occurs within the expanding clone of the first successful type-1 mutation (left panel of Figure 4). The initiation time σ_2 is exponential and does not depend on the spatial dimension.
- Regime 2: (R2) For $\Gamma \in (10, 100)$, the first successful type-2 mutation occurs within one of several successful type-1 clones (middle panel in Figure 4). The initiation time is no longer exponential and depends explicitly upon the spatial dimension.
- Regime 3 (R3): When $\Gamma > 1000$, the first successful type-2 mutation occurs after many successful type-1 mutations have occurred (right panel of Figure 4). The first successful type-2 can arise from *either* a successful *or* an unsuccessful type-1 family; the initiation time represents a mixture distribution of these two events.
- Note that for $\Gamma \in [1, 10]$ and $\Gamma \in [100, 1000]$ we say that we are in borderline regimes R1/R2 and R2/R3 respectively.

We refer the reader to [26] for mathematical details of these statements. Note that these 'regimes' can be thought of as labels highlighting distinct types of initiation behaviors that arise as Γ changes. In fact the system behavior continuously varies through the parameter space, and borderline cases between these regimes do exist. Figure 5 shows how the distribution of the waiting time σ_2 varies with changing number of cells N in $d = 2$. We note that as N increases, the waiting time distribution shifts to the left and initiation occurs earlier. By comparing Figures 4 and 5 we see that early initiation times are associated with a diffuse premalignant field with a large number of independent lesions, whereas late initiation times are associated with a single premalignant field harboring the initiating tumor cell.

To briefly summarize, we have described first a microscopic model of cellular division, mutation and death within a regularly structured epithelial tissue. Analysis of the fine-scale dynamics of this model leads to a more tractable hybrid mesoscopic model which approximates the microscopic model. In the next section, we analyze this mesoscopic model to study the characteristics and extent of premalignant fields at the stochastic time of cancer initiation or diagnosis. In the analyses throughout, we will consider parameter ranges spanning all three regimes of initiation behavior; however, for simplicity in regime

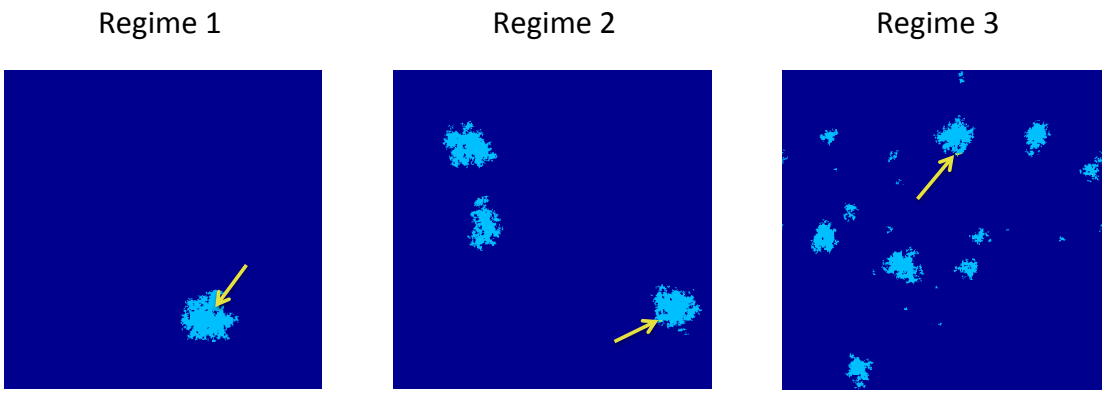

Figure 4: **The three dynamic regimes.** *Regime 1:* first successful type-2 cell (arrow) arises in the first premalignant clone, $\Gamma = 0.055$. *Regime 2:* several premalignant clones are present at the time of the first successful type-2 cell, $\Gamma = 54.47$. *Regime 3:* a large number of small premalignant clones are present by the time of the first successful type-2 cell, $\Gamma = 5.45 \times 10^4$. Simulations obtained with parameter values as in Figure 5.

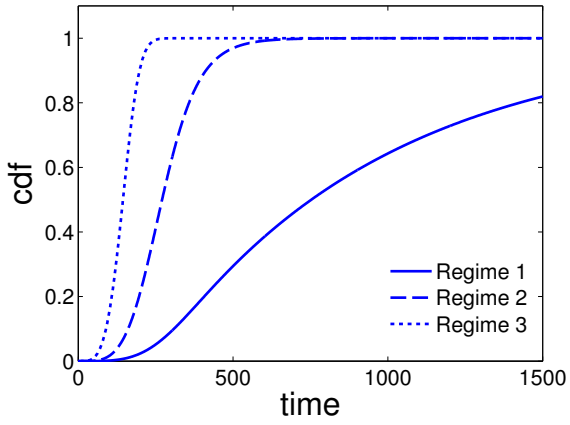

Figure 5: **Waiting time until first successful type-2.** Cumulative distribution function (cdf) of σ_2, the waiting time until the first successful type-2 mutation, for increasing N (see (4)). Regime 1: $u_1 = 7.5 \cdot 10^{-8}$, Regime 2: $u_1 = 7.5 \cdot 10^{-7}$, Regime 3: $u_1 = 7.5 \cdot 10^{-6}$. All other parameters are fixed: $d = 2$, $N = 2 \cdot 10^5$, $s_1 = s_2 = 0.1$, $u_2 = 2 \cdot 10^{-5}$, $c_2(s_1) = 0.16$.

3 we will restrict ourselves to the range of parameter space in which successful type-2 mutations arise from successful type-1 mutations (i.e. that do not later die out). The behavior in the final remaining portion of the parameter space in regime 3 will be the subject of further work.

3 Characterizing the premalignant field

The time between cancer initiation and diagnosis, which we label here as T_D, is a subject of great interest, see e.g. [33] for a review. In general, T_D is itself a random variable and may depend on the natural history of the disease until initiation. However, if we assume that T_D is independent of σ_2, then we can characterize the premalignant field at time of diagnosis, $\sigma_2 + T_D$, by means of the field characterization at time σ_2, together with the distribution of the delay time T_D. For this reason, even though the clinically relevant time is $\sigma_2 + T_D$, we focus here on characterizing the field at σ_2. Note that mathematically, this requires us to condition our analyses upon observing σ_2 at some time t, i.e. condition upon the event $\{\sigma_2 = t\}$.

The starting time of the model ($t = 0$) is assumed to be at the end of tissue development and the start of the tissue renewal phase. However for some tissues it is difficult to estimate this time, and thus it may be difficult to ascertain the system time t at the time σ_2. In such cases, it is simple to adapt our analyses to this scenario and treat σ_2 as an unobservable quantity, by removing the conditioning on $\{\sigma_2 = t\}$ and integrating of our results against the density of σ_2, which is given by (see (24) in section 7.1 for derivation)

$$\lambda e^{t\lambda(\phi(t)-1)} \left(1 - e^{-\theta t^{d+1}}\right), \tag{4}$$

where

$$\phi(t) \equiv \frac{1}{t} \int_0^t \exp\left(-\theta r^{d+1}\right) dr. \tag{5}$$

The constants in (4) and (5) are the arrival rate of successful type-1 mutations

$$\lambda \equiv N u_1 \bar{s}_1, \tag{6}$$

and

$$\theta \equiv \frac{u_2 \bar{s}_2 \gamma_d c_d^d(s_1)}{d+1}, \tag{7}$$

where we used the notation $\bar{s}_i = s_i/(1 + s_i)$.

3.1 Size of the local field at initation

We are first interested in characterizing the size of the local field, i.e. the region of the premalignant type-1 clone that gives rise to the first successful type-2 clone (see Figure 6). Following the nomenclature of [34], we note the distinction between two different types of recurrent tumors: if the recurrence arises from a transformed cell in the premalignant field that gave rise to the primary tumor, the recurrence is called a *second field tumor*, see Figure 6A. On the other hand, if the recurrence arises from a premalignant field that is clonally unrelated to the primary malignancy, it is called a *second primary tumor*, see Figure 6B. These two types of recurrent tumors vary in terms of their degree of clonal relatedness to the primary tumor, and this may have some implications for treatment strategies in primary vs. recurrent tumors.

We define now $R_l(t)$ to be the radius of the local field at time t, and $X_l(t)$ its corresponding area ($X_l = \gamma_d R_l^d$). Note that we will use the terminology 'area' to describe clone sizes in all dimensions, and reserve the use of the term 'volume' for space-time quantities. In the following, we are interested in determining the distributions of these two quantities at time σ_2, conditioned on the event $\{\sigma_2 = t\}$. In other words, we are looking for the distributions of $(R_l(\sigma_2)|\sigma_2 = t)$ and $(X_l(\sigma_2)|\sigma_2 = t)$, respectively.

At any given time, each clone produces initiating mutations at a rate proportional to its area. Hence the probability that clone i (born at time T_i) gives rise to the initiating mutation at time t is given by the ratio of clone i's own area,

$$X_i(t) \equiv \gamma_d c_d^d(s_1)(t - T_i)^d,$$

divided by the total area of type-1 clones present. In other words, the size distribution of the initiating clone is given by the distribution of a *size-biased pick* from the different clones present at the time the initiated mutation arises.

Definition 3.1 (Size-biased pick)**.** *Let $L_1, \ldots, L_n$ be a family of n random variables. A size-biased pick from $L_1, \ldots, L_n$ is defined as a random variable $L_{[1]}$ with conditional probability distribution*

$$P(L_{[1]} = L_i | L_1, \ldots, L_n) = L_i / \sum_{j=1}^{n} L_j.$$

The following theorem is the main result of this section and characterizes the size-distribution of the local field at the time of initiation. This is recognized as a size-biased pick from the clones present at time t, conditioned on the event $\{\sigma_2 = t\}$.

Theorem 3.2. *The distribution of the area of the local field at time σ_2, conditioned on $\{\sigma_2 \in dt\}$, is given by*

$$\hat{P}\left(X_l(\sigma_2) \in dx\right) = \hat{P}\left(X_{[1]} \in dx\right) = \frac{u_2 \bar{s}_2 x^{1/d}}{d\gamma_d^{1/d} c_d(s_1)(1 - e^{-\theta t^{d+1}})} \exp\left[\frac{-u_2 \bar{s}_2 x^{\frac{d+1}{d}}}{(d+1)\gamma_d^{1/d} c_d(s_1)}\right], \tag{8}$$

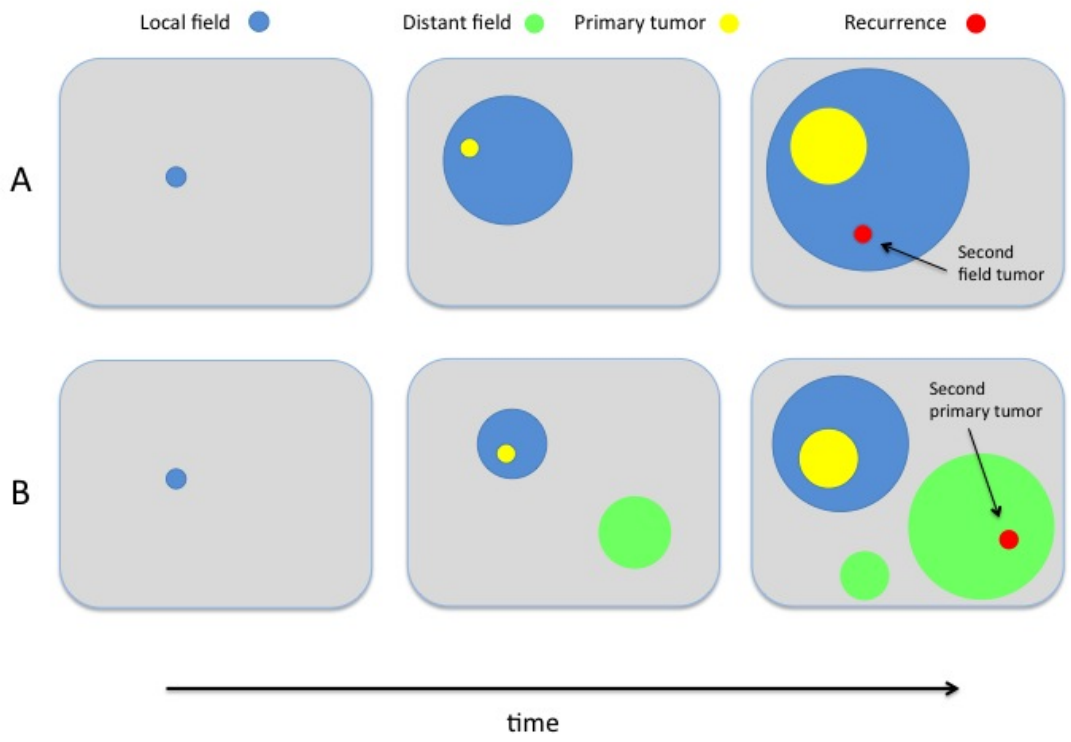

Figure 6: **Local and distant recurrences.** Local (blue) and distant (green) premalignant fields give rise to second field tumors and second primary tumors (both red), respectively. In scenario **A**, there is only one premalignant field (the local field) present at time of cancer initiation (middle panel), and the recurrence occurs inside the local field. In scenario **B**, two unrelated precancerous fields are present at time of initiation (middle panel), and the recurrence may occur as a second primary tumor in the distant field.

for $x \in [0, \gamma_d c_d^d(s_1) t^d]$.

The proof of this result is found in section 7.1, and the distribution of the local field radius follows easily as

$$\hat{P}(R_l(\sigma_2) \in dr) = \frac{u_2 \bar{s}_2 \, \gamma_d \, r^d}{c_d(s_1)(1 - e^{-\theta t^{d+1}})} \exp\left[-\frac{u_2 \bar{s}_2 \gamma_d r^{d+1}}{c_d(s_1)(d+1)} \right], \tag{9}$$

for $r \in [0, c_d(s_1)t]$.

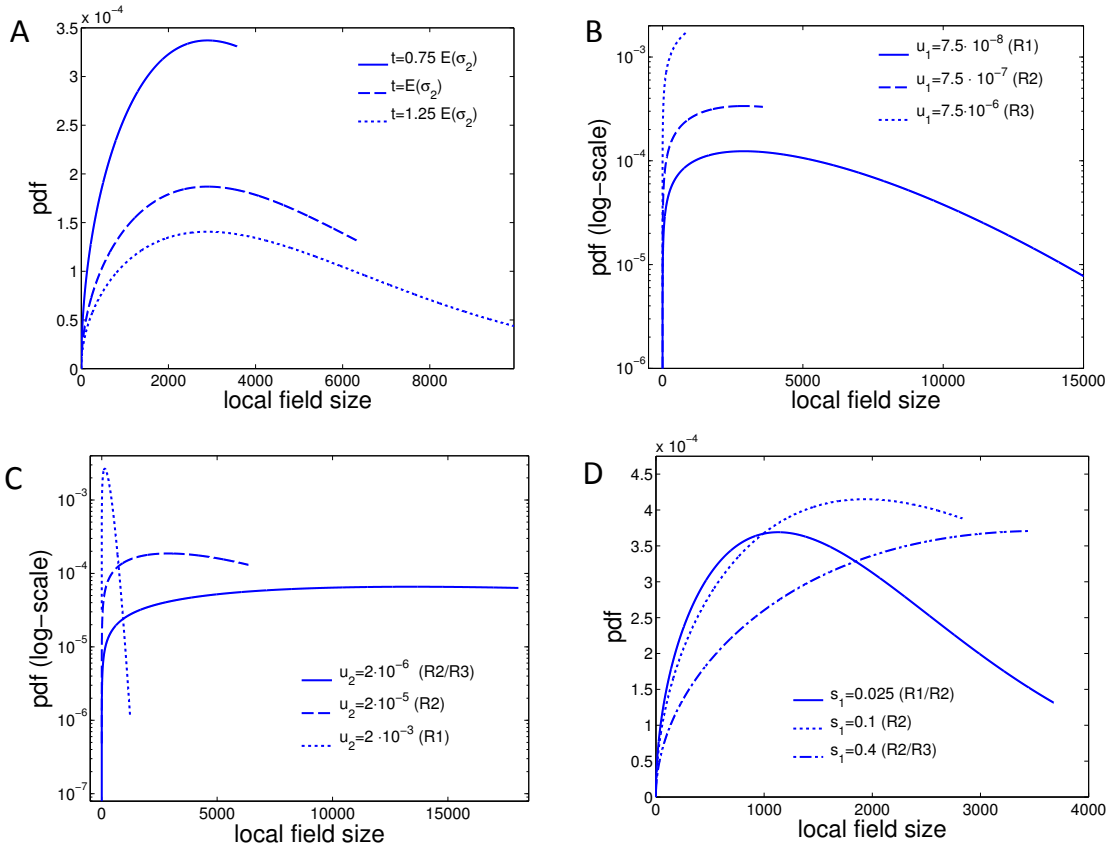

Figure 7: **Size-distribution of local field.** The size-distribution (8) of the local field is shown for different scenarios, corresponding to different Γ-values and regimes R1, R2 and R3 as explained in Section 2.4. **A** For varying arrival times t; **B** for varying type-1 mutation rates u_1; **C** for varying type-2 mutation rates u_2; (**D**) for varying type-1 fitness advantages s_1. The non-varying parameters are held constant at $d = 2$, $N = 2 \cdot 10^5$, $u_1 = 7.5 \cdot 10^{-7}$, $u_2 = 2 \cdot 10^{-5}$, $s_1 = s_2 = 0.1$ and $c_2(s_1) = 0.16$.

Note that the distribution of the local field size (8) depends on the rate of successful mutations $u_2\bar{s}_2$ and the growth rate $c_d(s_1)$, but is independent of λ, the arrival rate of type-1 mutations. In Figure 7A, we show how the distribution of the local field area (8) changes with arrival time of the first successful type-2 clone. As expected, the support of the distribution increases with increasing initiation time, and hence the likelihood of having a large local field increases substantially. This suggests that that tumors appearing later have a higher recurrence probability if only the malignant portion is removed during surgery. The finite support of each probability density function reflects the fact that there is a hard upper bound on the size of a premalignant field at finite time t in the system.

In Figure 7B,C we illustrate the sensitivity of the size-distribution of the local field to varying mutation rates u_1 and u_2, conditioned on observing initiation at the expected time $t = E(\sigma_2)$. The mutation rates are tuned to vary across parameter Regimes 1, 2, and 3 as described in the previous section. Observe that for lower mutation rates, the local field size varies widely (and sometimes close to uniformly) over a large range of values, while elevated mutation rates in both cases signify smaller local fields. For the u_1 rate (Figure 7B), an intuitive explanation for this behavior is that as the mutation rate increases, the system moves towards regimes 2 and 3, in which the premalignant field is comprised of an increasing number of independent type-1 patches. With more type-1 patches present, the space-time volume of type-1 cells that can give rise to the first successful type-2 cell increases faster, and hence the size of the patch that eventually gives rise to the first type-2 decreases accordingly. For u_2 (Figure 7C) on the other hand, an increase in the mutation rate signifies a move towards regime 1: fewer type-1 clones are required to produce the first successful type-2, and the size of the type-1 field that yields the first type-2 decreases with increasing u_2. Another observation to note is that the local field size varies across the same range of orders of magnitude as the mutation rates. This suggests for example, that carcinogen exposure or environmental causes changing mutation rates by one order of magnitude could result in predicted field sizes impacted similarly by an order of magnitude.

Finally, we demonstrate the sensitivity of the local field size to the selective advantage s of mutant cells, see Figure 7D. For a small fitness gain of $s = 0.025$, the distribution is peaked at lower field sizes, but as s increases the field size distribution shifts to the right. High fitness gains are usually associated with an aggressive tumor phenotype, and Figure 7D suggests that such tumors may also be associated with large surrounding premalignant fields and thus higher recurrence risks.

3.2 Size of the distant field at initiation

Next we are interested in analyzing the size distribution of the distant field at initiation, which is comprised of premalignant clones that are clonally unrelated to the tumor. Define the vector of areas of the distant premalignant lesions at time t to be $\bar{X}_d(t)$. This vector holds the areas of all premalignant clones except for the local field clone from which the tumor arises. Mathematically speaking, the goal of this section is to characterize the law of

$\bar{X}_d(\sigma_2)$ conditioned on the event $\{\sigma_2 = t\}$. Before stating the main result some additional notation is needed. First, define the mapping $\alpha_j(i)$ as follows:

$$\alpha_j(i) = \begin{cases} i, & \text{if } j > i \\ i+1, & \text{if } j \leq i. \end{cases}$$

Then, we define the random variable $\tilde{X}_i \equiv X_{\alpha(i)}$, where

$$\alpha(i) \equiv \sum_{j=1}^{M(t)} \alpha_j(i) 1_{\{X_{[1]} = X_j\}}.$$

Note that using this definition, $(\tilde{X}_1, \ldots, \tilde{X}_{M(\sigma_2)-1})$ represents the vector of sizes of the clones present at time σ_2, omitting the entry corresponding to the size-biased pick $X_{[1]}$ which represents the local field. In other words, the distribution of $\bar{X}_d(\sigma_2)$ is the joint distribution of $(\tilde{X}_1, \ldots, \tilde{X}_{M(\sigma_2)-1})$, which characterizes the size distribution of the clones in the distant field at time σ_2. We obtain the following result (see section 7.2 for the proof).

Theorem 3.3. *The size-distribution of the distant field clones at time σ_2 of the first successful type-2 mutation, conditioned on $\{\sigma_2 = t\}$, is given by*

$$\begin{aligned} \mathcal{L}(\bar{X}_d| \in dt) &=_d \hat{P}(\tilde{X}_1 \in dx_1, \ldots, \tilde{X}_{M(t)-1} \in dx_{M(t)-1}) \\ &= \frac{1}{1 - e^{-\lambda t \phi(t)}} \sum_{m=1}^{\infty} \frac{(\lambda \phi(t) t)^m e^{-\lambda \phi(t) t}}{m!} \prod_{i=1}^{m-1} g_t(x_i), \end{aligned}$$

where $g_t(x)$ is defined in (26).

Of note, from Theorem 3.3 and Corollary 3.5 below, we see that

$$\mathcal{L}(\bar{X}_d | \sigma_2 = t, M(t) = m) =_d \hat{P}(\tilde{X}_1 \in dx_1, \ldots, \tilde{X}_{m-1} \in dx_{m-1}) = \prod_{i=1}^{m-1} g_t(x_i).$$

Figure 8 shows how the probability density function of the total distant field size (i.e. the sum of all distant field patches) changes with increasing mutation rate u_1. For a comparison to the local field size distribution at the same parameter values, we refer to Figure 7B. We note that in regimes 1 and 2 the total distant field size is on the same order of magnitude as the local field size, but in regime three the distant field size is significantly larger than the size of the local field. As will be investigated in more detail below, this suggests that secondary tumor recurrences for cancer types in regime 3 are much more likely to stem from the distant field, and thus are more likely to be clonally unrelated to the primary tumor.

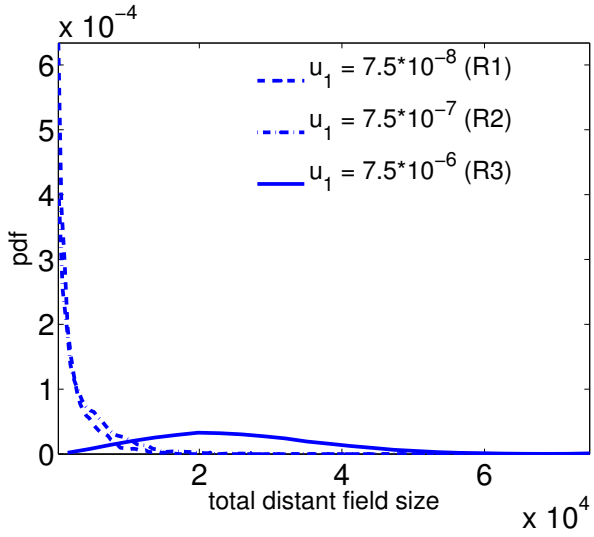

Figure 8: The distribution of the total size of the distant field is shown for different scenarios, corresponding to the three regimes R1, R2 and R3 illustrated in Figure 4 for varying type-1 mutation rates u_1. The non-varying parameters are held constant at $d = 2$, $N = 2 \cdot 10^5$, $u_2 = 2 \cdot 10^{-5}$, $s_1 = s_2 = 0.1$ and $c_2(s_1) = 0.16$.

3.3 Number of field patches: evolution until initiation

We next analyze the total number of premalignant lesions over time until tumor initiation. In particular, the following result holds (see section 7.3 for the proof).

Proposition 3.4. *Conditioned on $\{\sigma_2 = t\}$, we have that for all $\zeta \leq t$, the number of field patches is distributed as a mixture of a Poisson and a shifted Poisson random variable. In particular,*

$$P(M(\zeta) = m|\sigma_2 = t) = p_1(t,\zeta)\,\frac{\lambda^m\,[t\phi(t) - (t-\zeta)\phi(t-\zeta)]^m}{(m)!}e^{-\lambda[t\phi(t)-(t-\zeta)\phi(t-\zeta)]} + p_2(t,\zeta)\,\frac{\lambda^{m-1}\,[t\phi(t) - (t-\zeta)\phi(t-\zeta)]^{m-1}}{(m-1)!}e^{-\lambda[t\phi(t)-(t-\zeta)\phi(t-\zeta)]},$$

where $p_1(t,\zeta) + p_2(t,\zeta) = 1$ and $p_1(t,\zeta) = (1 - e^{-\theta(t-\zeta)^{d+1}})/(1 - e^{-\theta t^{d+1}})$. In particular,

$$E(M(\zeta)|\sigma_2 = t) = \lambda\,[t\phi(t) - (t-\zeta)\phi(t-\zeta)] + p_2(t,\zeta).$$

It is interesting to observe that as $\zeta \to t$ we see that $p_1(t,\zeta) \to 0$, therefore as ζ gets closer to time t the process looks more like a shifted Poisson. This is stated in the corollary below.

Corollary 3.5.

$$\hat{P}(M(t)=m)=\frac{(\lambda\, t\, \phi(t))^{m-1}}{(m-1)!}e^{-t\lambda\phi(t)}, \quad m\geq 1, \tag{10}$$

and $\hat{P}(M(t)=m)=0$. *In particular,*

$$\hat{E}(M(t))=1+E(M(t)|\sigma_2>t)=1+\lambda t\phi(t), \tag{11}$$

where $E(M(t)|\sigma_2>t)$ *is discussed in Lemma 7.2.*

Using Proposition 3.4, we can study the expected number of field patches of a certain size over time. Figure 9 shows the temporal dynamics of clone-size distribution in each regime. In regime 1 the expected number of small clones peaks and then declines as larger clones begin to dominate (consistent with the notion that a single premalignant clone exists prior to initiation), whereas in regimes 2 and 3 we see longer coexistence of large and small clones over time.

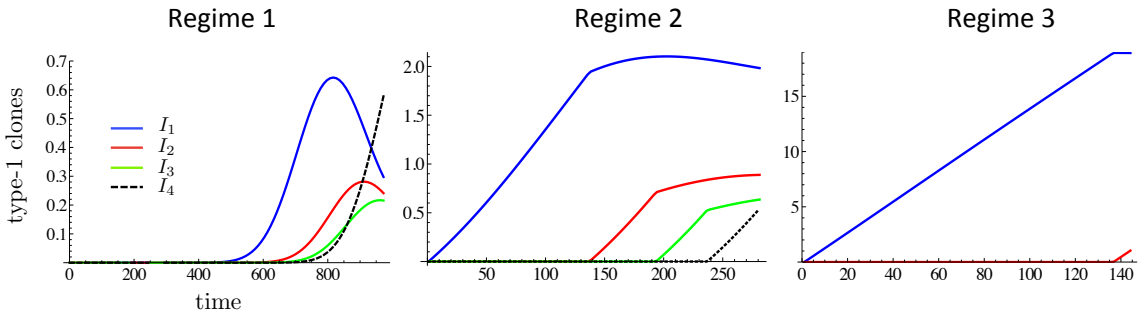

Figure 9: **Dynamic clone-size distribution.** For each of the three regimes in Figure 5, the expected number of type-1 clones of sizes comprised in the corresponding intervals I_j are shown as functions of time up to $E(\sigma_2)$ (expectations are conditioned on $\{t=E(\sigma_2)\}$). The intervals are defined as $I_1=[0,1500)$, $I_2=[1500,3000)$, $I_3=[3000,4500)$ and $I_4=[4500,+\infty)$. Parameter values as in Figure 5.

Finally, we would like to point out that the result in Proposition 3.4 can be extended to a result about the entire process $\{M(r):0\leq r\leq t\}$ conditioned on $\sigma_2=t$. The details are provided in section 7.4.

4 Recurrence predictions

Tumor recurrence due to field cancerization poses a substantial clinical problem in many epithelial cancers [3]. We next aim to use the results of the previous section to develop a methodology for assessing the risk of tumor recurrence (as well as the likely type of tumor recurrence) after surgical removal of the primary tumor.

4.1 Local vs. distant field recurrence?

As discussed above, a recurring tumor can either arise in the same premalignant field (a second field tumor), or it can arise in a clonally unrelated field (second primary tumor). In this section we characterize the recurrence time distribution for each of these secondary tumor types, and study how the relative likelihood of local vs. distant recurrence depends upon parameters of the tissue and cancer type.

To this end, we first study the recurrence time distribution for second field tumors, which arise from the local premalignant field. Denote the second field recurrence time by T_R^f, measured in time units τ starting from $\tau = 0$ at time σ_2. The time is reset at the tumor initiation time σ_2, rather than the tumor resection time $\sigma_2 + T_D$, to accommodate the possibility that a recurrence occurs *prior* to detection of the primary tumor. Thus if recurrence occurs at some time $\tau < T_D$, then a secondary tumor already exists at the time of diagnosis of the primary tumor (but may be too small to be detectable). We assume that the primary tumor node is completely resected once it becomes detectable at time T_D, leaving the surrounding field intact (i.e. there are no excision margins).

At time σ_2 a successful type-2 cell arises from a premalignant clone of radius $R_l(\sigma_2)$, whose distribution is characterized in (9). If $R_l(\sigma_2) = r$, the incidence rate of successful type 2 mutations within this field is given by

$$\eta(r, \tau) \equiv u_2 \bar{s}_2 \gamma_d \left[(r + c_d(s_1)\tau)^d - c_d^d(s_2) \left(\tau \wedge T_D\right)^d \right], \tag{12}$$

where $c_d(s_2)$ is the rate of expansion of the malignant cells into the type-1 field. The proof of the following result can be found in section 7.5.

Corollary 4.1. *The probability of a second field tumor having formed before time τ (measured from σ_2), conditioned on $\{\sigma_2 = t\}$, is given by*

$$\hat{P}(T_R^f < \tau) = 1 - \frac{\gamma_d u_2 \bar{s}_2}{c_d(s_1)(1 - e^{-\theta t^{d+1}})} \int_0^{c_d(s_1)t} r^d \exp\left[-\frac{u_s \bar{s}_2 \gamma_d}{c_d(s_1)(d+1)} r^{d+1} - \int_0^\tau \eta(r, s) ds \right] dr.$$

In particular, $\hat{P}(T_R^f < T_D)$ is the probability that smaller, possibly undetectable second field tumors exist at the time of diagnosis.

In Figure 10A the cumulative distribution function of T_R^f as calculated in Corollary 4.1 is shown, for varying values of type-2 mutation rates u_2. As one might expect, higher mutation rates yield a decreased time to recurrence (the curves shift to the left for increasing u_2). However, considering that the size of the premalignant field at initiation of the primary tumor is inversely proportional to the mutation rate u_2, see Figure 10B, the decrease in time to recurrence is *a priori* not obvious: a bigger precancer field increases the chance of fast recurrence. This example illustrates how a quantitative model enables us to assess the relative importance of competing aspects of the system - in this case, the impact of larger premalignant field versus higher mutation rates on recurrence likelihood.

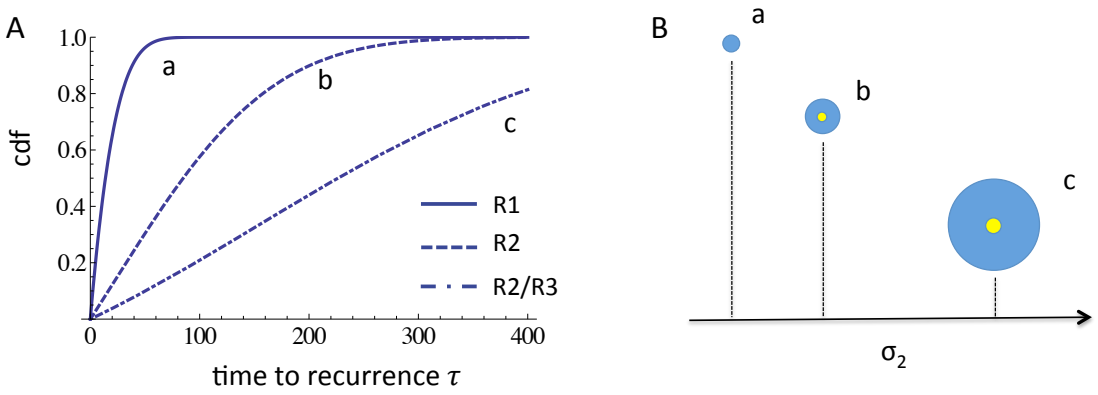

Figure 10: **Time to local recurrence. A** The cumulative distribution function of the time to recurrence of a second field tumor is shown for three different scenarios, corresponding to $u_2 = 2 \cdot 10^{-3}$ (Regime 1), $u_2 = 2 \cdot 10^{-5}$ (Regime 2) and $u_2 = 2 \cdot 10^{-3}$ (Regime 2/3), respectively. The remaining parameters are $d = 2$, $N = 2{\cdot}10^5$, $u_1 = 7.5{\cdot}10^{-7}$, $s_1 = s_2 = 0.1$, $t = E(\sigma_2)$. **B** Schematic of the relative initiation times of the primary tumor (yellow) and sizes of the local fields (blue), for the three scenarios in panel A. The numerical values for expected initiation time and local field size are: **(a)** $\mathbb{E}(\sigma_2) = 123$, $\hat{E}(R_l) = 8$; **(b)** $\mathbb{E}(\sigma_2) = 281$, $\hat{E}(R_l) = 31$; **(c)** $\mathbb{E}(\sigma_2) = 474$, $\hat{E}(R_l) = 55$.

If the recurrence does not take place in the local field giving rise to the first successful type-2 clone, then it either arises from one of the type-1 clones already present at time of initiation (i.e. the distant field), or it arises in a type-1 clone formed after initiation. In the latter case, the waiting time is again distributed as σ_2, and hence we focus here on the distribution of the waiting time T_R^p, defined as the time from σ_2 until a second primary tumor arises from the distant field already existing at σ_2. We have the following result, proved in section 7.6.

Corollary 4.2. *The probability that the distant field at the time of initiation gives rise to a second primary tumor by time τ (measured from σ_2), conditioned on $\{\sigma_2 = t\}$, is given by*

$$P(T_R^p > \tau | \sigma_2 = t) = \exp\left[-\lambda t \phi(t)\left(1 - d\gamma_d \Phi(\tau, t)\right)\right]$$

where

$$\Phi(\tau, t) = \int_0^\infty \exp\left(-\int_0^\tau \eta\left(r, s\right) ds\right) r^{d-1} g_t(r_i^d \gamma_d) dr,$$

and g_t is defined in (26).

Thanks to the results in this section, it is now possible to evaluate the probability of local versus distant tumor recurrences in each parameter regime. Corollary 4.1 explicitly provides the probability density function $\hat{P}(T_R^f \in d\tau)$, which is the probability that a second field tumor arises at time τ from the same field that gave rise to the primary tumor. To obtain the corresponding probability density function for recurrence as a second primary tumor, we have to consider recurrences due to distant field lesions that have arisen before and after σ_2. While Corollary 4.2 characterizes the recurrence risk due to distant lesions already present at initiation, the time to a successful second primary tumor from a distant field not yet present at initiation is distributed as σ_2, see (4). Therefore, the distribution of interest is that of $\tilde{T}_R^p = \min\{T_R^p, \sigma_2\}$, which is the time of the first distant recurrence event.

In Figure 11 we study how the comparison between the probability density functions of T_R^f (second field tumor, local) and $\tilde{T}_R^p$ (second primary tumor, distant) varies in regimes 1, 2 and 3. The likelihood of local vs. distant recurrences depends strongly upon both the timing and parameter regime of the system In regime 1, local recurrence is significantly more likely overall, but at late times the probability of distant recurrences is slightly higher than for local recurrences. In contrast, in regimes 2 and 3 the overall probability of local and distant recurrences are comparable. However, in regime 2, at early times distant field recurrences are more likely, whereas the opposite is true at later times. The same observation, but even more pronounced, holds in regime 3.

5 Conclusions and outlook

In this study we performed a quantitative analysis of the cancer field effect by means of a spatial stochastic model of cancer initiation, which had previously been introduced in [26]. Using this model, we studied the characteristics of premalignant fields at the time of tumor initiation. In particular, we derived the size-distributions of the local field (the premalignant lesion that gives rise to the tumor) and the distant field (the premalignant lesions that are unrelated to the primary tumor). We also investigated how the extent and geometry of these fields depend upon Γ, a key combination of parameters of the tissue and genetic pathway leading to cancer. We calculated the dynamic clone size distribution at times leading up to initiation, and derived the probability density functions of local and distant recurrence times. Finally, we compared the relative likelihood of second field versus second primary tumors, and demonstrated how the clonal relatedness between primary and recurrent tumors depends explicitly upon tissue and cancer type parameters.

Using an example set of biologically realistic parameters in two space dimensions (which is appropriate for describing the cancer initiation process in the basal layer of a stratified epithelium), we found that lower mutation rates (such as in regime 1) were associated with larger local field sizes, whereas higher mutation rates (regimes 2 and 3) led to smaller local fields. We also found that higher mutation rates resulted in larger distant fields, while more

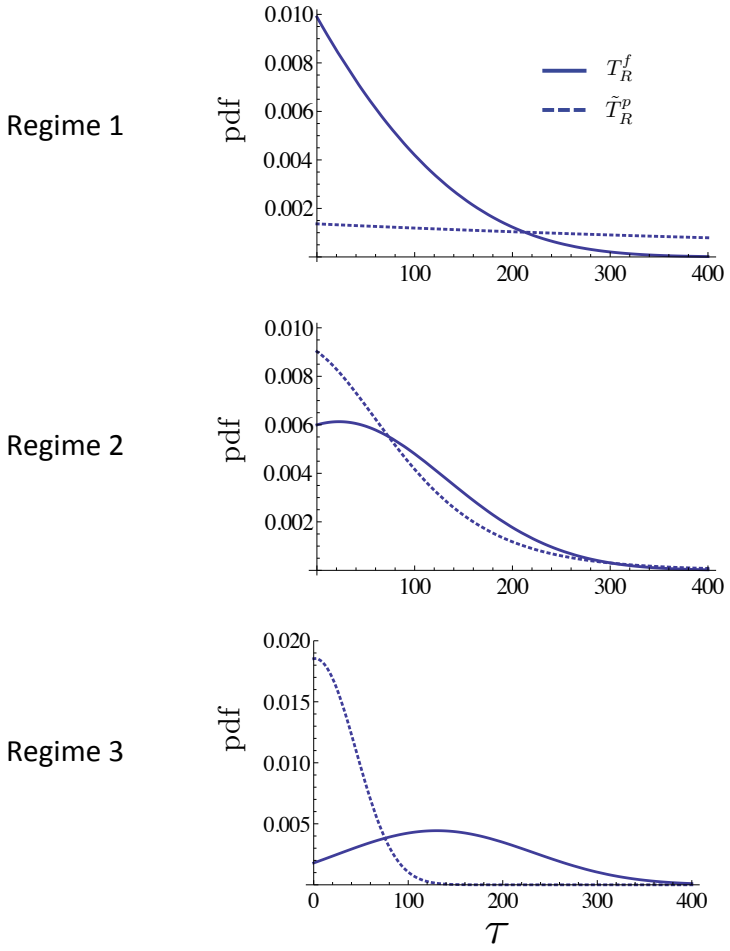

Figure 11: **Local vs. distant recurrence. A** For each of the three regimes in Figure 5, we show: the distribution of time to local recurrence $\hat{P}(T_R^f \in d\tau)$, and the distribution of time to distant recurrence $\hat{P}(\tilde{T}_R^p \in d\tau)$. The distribution of T_R^f is given in Corollary 4.1 and we set $\tilde{T}_R^p = \min\{T_R^p, \sigma_2\}$ to account both for contributions from type-1 clones already existing at σ_2 as well as contributions from type-1 clones born after σ_2 (for which time to recurrence is distributed as σ_2). Expected times to recurrence: $\hat{E}(T_R^f) = 81$ and $\hat{E}(\tilde{T}_R^p) = 733$ (Regime 1); $\hat{E}(T_R^f) = 98$ and $\hat{E}(\tilde{T}_R^p) = 86$ (Regime 2); $\hat{E}(T_R^f) = 149$ and $\hat{E}(\tilde{T}_R^p) = 34$ (Regime 3). The parameter values are as in Figure 5.

aggressive cancers (high selective advantage) led to larger local fields at diagnosis. Finally, we investigated the risk of recurrence after surgical resection of the malignant portion, and

found that for low mutation rates (regime 1), local recurrence is much more likely, whereas for larger mutation rates (regimes 2 and 3), the overall probability of local and distant recurrences are comparable. However, in regimes 2 and 3, early recurrences are more likely to be a second primary tumor, whereas the late recurrences are more likely to be second field tumors.

One important limitation of our approach is that the model captures a specific sequence of genetic alterations with specified u_i and s_i, and does currently not allow for permutations of genetic events and divergent pathways. Nevertheless, our model may provide a useful framework for comparing different biological hypotheses and disentangling divergent genetic pathways among cancer subtypes. In particular, it enables us to predict differences in observable dynamics such as initiation times and prognoses between different molecular models. Such an approach could help elucidating the sequence of genetic events during carcinogenesis, and will be the subject of future work. Another limitation of our framework is that we have assumed a static, uniform microenvironment within the tissue. The local microenvironment is in reality determined by a variety of time- and space-dependent factors such as glucose, oxygen, growth factors, drugs and cytokine concentrations. In addition to impacting the growth and mutation rates of cells within the tissue, the local microenvironment is increasingly being recognized as playing an important role in carcinogenesis through stromal signaling.

As mentioned before, field cancerization poses various clinical challenges, especially in the case of head and neck, where multifocal primary cancers as well as recurrences are common [35]. In particular, the optimal size of excision margins and assessment of the recurrence risk after surgery are largely unsolved problems arising in everyday clinical practice. In a forthcoming study, we will discuss how our analysis can be used to address some of the most pertinent clinical questions in head and neck cancer care.

In summary, the analyses performed in this work contribute towards a quantitative understanding of how organ-specific physiological parameters and pathway-specific parameters influence the process of field cancerization and the associated risk of recurrence. We demonstrate that tumor recurrence dynamics and premalignant field characteristics are strongly dependent upon these parameters, which vary across different tissue and cancer types. Once properly calibrated for a specific tissue and cancer type, the proposed methodology can potentially be used to provide insights into key prognostic factors such as risk of multifocal lesions and tumor recurrence, surveillance guidelines, and treatment design. For example, we are able to assess the likelihood and timing of local versus distant recurrences after surgical resection. Since this distinction provides information on the level of clonal relatedness between primary and recurrent tumors, the model predictions may provide insights into whether treatment strategies effective for primary tumors will be useful for recurrent tumors in particular cancer types. In addition, our methodology can be utilized to assess the relative benefits of surgical excision margins, and to help determine the minimal margins necessary to prevent recurrence in each tissue type.

6 Acknowledgements

We thank Rick Durrett for insightful discussions on this project as well as his useful suggestions on the manuscript.

7 Appendix: Proofs

7.1 Proof of Theorem 3.2

To prove Theorem 3.2, we first need a few new definitions and preliminary results. Define $V(t)$ to be the random total space-time volume covered by successful type-1 families until time t,

$$V(t) = \sum_{i=1}^{M(t)} \gamma_d c_d^d(s_1) \frac{(t - T_i)^{d+1}}{d+1}, \tag{13}$$

where T_i represents the arrival time of the i-th family, and $M(t)$ is the total number of successful arrivals by time t, which is a Poisson process with rate λ. Let $V_{\mathcal{E}_t}$ represent the space-time volume conditioned on the event

$$\mathcal{E}_t(t_1, \ldots, t_m) \equiv \{M(t) = m, T_1 \in dt_1, \ldots T_m \in dt_m\},$$

where $0 < t_1 < \cdots < t_m < t$. In other words,

$$V_{\mathcal{E}_t} \equiv \frac{\gamma_d c_d^d(s_1)}{d+1} \sum_{i=1}^{m} (t - t_i)^{d+1}. \tag{14}$$

For ease of notation we replace $V_{\mathcal{E}_t(t_1,\ldots,t_n)}$ with the more compact version $V_{\mathcal{E}_t}$. Since $E[V(t)] = E[E[V(t)|M(t)]]$ and the conditioned process is a compound Poisson process, we obtain that

$$E[V(t)] = \sum_{m=0}^{\infty} P(M(t) = m) \frac{m \gamma_d c_d^d(s_1)}{d+1} E[(t - T_i)^{d+1}] = \lambda \gamma_d c_d^d(s_1) \frac{t^{d+2}}{(d+2)(d+1)}.$$

Similarly, we define $A(t)$ to be the total area of clones covered by successful type-1 families at time t,

$$A(t) \equiv \sum_{i=1}^{M(t)} \gamma_d c_d^d(s_1)(t - T_i)^d, \tag{15}$$

and we define $A_{\mathcal{E}_t}$ to be this quantity conditioned on $\mathcal{E}_t(t_1, \ldots, t_m)$,

$$A_{\mathcal{E}_t} \equiv \sum_{i=1}^{m} \gamma_d c_d^d(s_1)(t - t_i)^d. \tag{16}$$

Note that

$$E[A(t)] = \sum_{m=0}^{\infty} P(M(t)=m) m\gamma_d c_d^d(s_1) E[(t-T_i)^d] = \lambda\gamma_d c_d^d(s_1)\frac{t^{d+1}}{d+1}. \tag{17}$$

By considering the space-time volume of type-1 clones we can calculate $P(\sigma_2 > t|\mathcal{E}_t(t_1,\ldots,t_m)$ and $P(\sigma_2 > t|M(t)=m)$. Combining these two formulas and using Bayes rule we get the following result for the joint distribution of the arrival times of successful type-1 mutations, conditioned on the total number of mutations by time t.

Lemma 7.1. *Conditioned on $\{\sigma_2 > t\}$ and $\{M(t)=m\}$, the arrival times of successful type-1 clones $(T_1,\ldots,T_m)$ are distributed as order statistics of iid random variables as follows:*

$$P(T_1 \in dt_1,\ldots,T_m \in dt_m|\sigma_2 > t, M(t)=m) = \frac{m!}{t^m\phi(t)^m}\prod_{i=1}^{m} e^{-\theta(t-t_i)^{d+1}}$$

where $0 < t_1 < \cdots < t_m < t$.

Proof. The arrival process of successful type-1 mutations is represented by $M(\cdot)$, which is a Poisson process with rate $\lambda = Nu_1s_1/(1+s_1)$ and arrival times $T_1, T_2, \ldots$. Then for any $t > 0$ and sequence $0 < t_1 < \cdots < t_m < t$ we have that

$$P(\mathcal{E}_t(t_1,\ldots,t_m)) = \lambda^m e^{-\lambda t}. \tag{18}$$

Since

$$P(\sigma_2 > t|\mathcal{E}_t(t_1,\ldots,t_m)) = \exp(-u_2\bar{s}_2 V_{\mathcal{E}_t}), \tag{19}$$

we find using Bayes' rule

$$P(\sigma_2 > t, \mathcal{E}_t(t_1,\ldots,t_m)) = \lambda^m e^{-\lambda t}\exp(-u_2\bar{s}_2 V_{\mathcal{E}_t}).$$

It follows then that

$$\begin{aligned}
&P(T_1 \in dt_1,\ldots,T_m \in dt_m|\sigma_2 > t, M(t)=m) = \frac{P(\sigma_2 > t, \mathcal{E}_t(t_1,\ldots,t_m))}{P(\sigma_2 > t|M(t)=m)P(M(t)=m)}\\
&= \frac{\lambda^m e^{-\lambda t}\exp\left(-u_2\bar{s}_2 V_{\mathcal{E}_t}\right)}{P(\sigma_2 > t|M(t)=m)e^{-\lambda t}(\lambda t)^m/m!}\\
&= \frac{m!}{t^m}\frac{\exp\left(-u_2\bar{s}_2 V_{\mathcal{E}_t}\right)}{\left(E\exp\left(-u_2\bar{s}_2\gamma_d c_d^d(s_1)(t-T)^{d+1}/(d+1)\right)\right)^m}\\
&= m!\prod_{i=1}^{m}\left(\frac{1}{t}\right)\frac{\exp\left(-u_2\bar{s}_2\gamma_d c_d^d(s_1)(t-t_i)^{d+1}/(d+1)\right)}{E\exp\left(-u_2\bar{s}_2\gamma_d c_d^d(s_1)(t-T)^{d+1}/(d+1)\right)},
\end{aligned}$$

where T is a uniform random variable on $[0,t]$. . □

The distribution in Lemma 7.1 is an exponential twist of the uniform distribution. Note that if the conditioning was placed on the set $\{\sigma_2 = t\}$ instead of $\{\sigma_2 > t\}$, then the conditional distribution would no longer have product form because of the term $\frac{d}{dt}V_{\mathcal{E}_t}$, and the arrival times would not be the order statistics from an iid collection of random variables.

Next, we show that the random variable $M(t)$ is Poisson if conditioned on $\{\sigma_2 > t\}$.

Lemma 7.2. *Conditioned on* $\{\sigma_2 > t\}$, $M(t) =_d Pois\left(\lambda t \phi(t)\right)$.

Proof. First we note that

$$\begin{aligned} P(\sigma_2 > t) &= \sum_{m=0}^{\infty} \frac{1}{m!} \int_{[0,t]^m} P(\sigma_2 > t | \mathcal{E}_t(t_1, \ldots, t_m)) P(\mathcal{E}_t(t_1, \ldots, t_m)) dt_1 \ldots dt_m \\ &= \sum_{m=0}^{\infty} \frac{1}{m!} \int_{[0,t]^m} \exp\left(-u_s \bar{s}_2 V_{\mathcal{E}_t}\right) \lambda^m e^{-\lambda t} dt_1 \ldots dt_m \\ &= \sum_{m=0}^{\infty} \frac{1}{m!} t^m \lambda^m \ e^{-\lambda t} \left(\frac{1}{t} \int_0^t \exp\left(-\frac{u_2 \bar{s}_2 \gamma_d c_d^d(s_1)(t-r)^{d+1}}{d+1} \right) dr \right)^m \\ &= \sum_{m=0}^{\infty} \frac{(t\lambda\phi(t))^m}{m!} e^{-\lambda t} = e^{t\lambda(\phi(t)-1)}. \end{aligned} \tag{20}$$

From this, we find using Bayes' rule

$$\begin{aligned} P(\mathcal{E}_t(t_1, \ldots, t_m) | \sigma_2 > t) =& \frac{P(\sigma_2 > t | \mathcal{E}_t(t_1, \ldots, t_m)\, P(\mathcal{E}_t(t_1, \ldots, t_m))}{P(\sigma_2 > t)} \\ =& \frac{\lambda^m e^{-\lambda t} \exp(u_s \bar{s}_2 V_{\mathcal{E}_t)}}{e^{t\lambda(\phi(t)-1)}}, \end{aligned} \tag{21}$$

and hence

$$\begin{aligned} P(M(t) = m | \sigma_2 > t) =& \frac{1}{m!} \int_{[0,t]^m} P(\mathcal{E}_t(t_1, \ldots, t_m) | \sigma_2 > t) dt_1 \ldots dt_m \\ =& e^{-\lambda t \phi(t))} \ \frac{(t\lambda\phi(t))^m}{m!}. \end{aligned}$$

. □

For subsequent considerations, it will be useful to define the two conditional probability measures $\hat{P}(\cdot) = P(\cdot | \sigma_2 = t)$ and $\tilde{P}(\cdot) = P(\cdot | \sigma_2 > t)$, and their corresponding expected values, $\hat{E}(\cdot) = E(\cdot | \sigma_2 = t)$ and $\tilde{E}(\cdot) = E(\cdot | \sigma_2 > t)$, respectively. In particular, we can compute the Radon-Nikodym derivative between these two measures.

Lemma 7.3. *The Radon-Nikodym derivative of $\hat{P}$ with respect to $\tilde{P}$ is given by*

$$\frac{d\hat{P}}{d\tilde{P}} = \frac{A_{\mathcal{E}_t} u_2 \bar{s}_2}{\lambda(1 - e^{-\theta t^{d+1}})}. \tag{22}$$

Proof. First, note that

$$P(\mathcal{E}_t(t_1, \ldots, t_m)|\sigma_2 = t) = \frac{P(\mathcal{E}_t(t_1, \ldots, t_m))P(\sigma_2 = t|\mathcal{E}_t(t_1, \ldots, t_m))}{P(\sigma_2 = t)}. \tag{23}$$

By differentiating (19) and (20) we obtain

$$P(\sigma_2 = t|\mathcal{E}_t(t_1, \ldots, t_m)) = u_2 \bar{s}_2 A_{\mathcal{E}_t} \exp(-u_2 \bar{s}_2 V_{\mathcal{E}_t})$$

and

$$P(\sigma_2 \in dt) = -\frac{d}{dt} e^{t\lambda(\phi(t)-1)} = \lambda \left(1 - e^{-\theta t^{d+1}}\right) e^{t\lambda(\phi(t)-1)}. \tag{24}$$

Hence (23) becomes

$$P(\mathcal{E}_t(t_1, \ldots, t_m)|\sigma_2 = t) = \lambda^m e^{-\lambda t} \frac{u_2 \bar{s}_2 A_{\mathcal{E}_t} \exp(-u_2 \bar{s}_2 V_{\mathcal{E}_t})}{\lambda e^{t\lambda(\phi(t)-1)}(1 - e^{-\theta t^{d+1}})},$$

and comparing this to (21) yields the desired result. □

Recall now that $M(t)$ is the number of successful type-1 mutations that have arrived by time t, and we denote their arrival times by $T_1, \ldots, T_{M(t)}$. At time t, the area of a clone created at time $r < t$ is $\gamma_d c_d^d(s_1)(t - r)^d$, and hence the area of the i-th clone at time t is given by the random variable

$$X_i(t) \equiv \gamma_d c_d^d(s_1)(t - T_i)^d.$$

Using the above results together with definition 3.1 of a size-biased pick we can now prove Theorem 3.2.

Proof of Theorem 3.2. Using basic properties of conditional expectations and Definition 3.1 we find

$$\begin{aligned}
\hat{P}(X_{[1]} \in dx) &= \hat{E}\left[\hat{P}(X_{[1]} \in dx|X_1, \ldots, X_{M(t)}, M(t))\right] \\
&= \hat{E}\left[\sum_{i=1}^{M(t)} \frac{X_i 1_{\{X_i \in dx\}}}{S_{M(t)}}\right] = \sum_{m=1}^{\infty} \hat{E}\left[\sum_{i=1}^{m} \frac{X_i 1_{\{X_i \in dx\}}}{S_m} 1_{\{M(t)=m\}}\right],
\end{aligned}$$

where $S_m = X_1 + \ldots + X_m$. Using the Radon-Nikodym derivative (22) we can rewrite this as

$$\begin{aligned}
&= \sum_{m=1}^{\infty} \tilde{E}\left[\frac{1_{\{M(t)=m\}} u_2 \bar{s}_2}{\lambda(1-e^{-\theta t^{d+1}})}\left(\sum_{i=1}^{m} \frac{X_i 1_{\{X_i \in dx\}}}{S_m}\right)\sum_{j=1}^{m} X_j\right] \\
&= \frac{u_2\bar{s}_2}{\lambda(1-e^{-\theta t^{d+1}})}\sum_{m=1}^{\infty}\tilde{E}\left[1_{\{M(t)=m\}}\sum_{i=1}^{m} x 1_{\{X_i\in dx\}}\right] \\
&= \frac{x u_2\bar{s}_2}{\lambda(1-e^{-\theta t^{d+1}})}\sum_{m=1}^{\infty} E\left[\sum_{i=1}^{m} 1_{\{X_i\in dx\}} | M(t)=m, \sigma_2 > t\right] P(M(t)=m|\sigma_2>t) \\
&= \frac{x u_2\bar{s}_2}{\lambda(1-e^{-\theta t^{d+1}})} P(X_1(t)\in dx | M(t)=m,\sigma_2>t) E\left[M(t)|\sigma_2>t\right],
\end{aligned} \tag{25}$$

where we have used the fact that $P(X_1(t) < x|M(t) = m, \sigma_2 > t)$ is independent of m, which we will show below. Using Lemma 7.1 and differentiating the cumulative distribution function

$$P(X_1(t) < x|M(t) = m, \sigma_2 > t) = P\left(T_1 > t - \left(\frac{x}{\gamma_d c_d^d(s_1)}\right)^{1/d} \Big| M(t) = m, \sigma_2 > t\right),$$

we determine that

$$P(X_1(t)\in dx|M(t)=m,\sigma_2>t) = \frac{x^{1/d-1}}{d\gamma_d^{1/d} c_d(s_1) t\phi(t)} \exp\left[\frac{-u_2\bar{s}_2 x^{\frac{d+1}{d}}}{(d+1)\gamma_d^{1/d} c_d(s_1)}\right] \equiv g_t(x) \tag{26}$$

for $x \in [0, \gamma_d c_d^d(s_1) t^d]$. Note that (26) is indeed independent of m. From Lemma 7.2 it follows that

$$E\left[M(t)|\sigma_2 > t\right] = \lambda t\phi(t),$$

and combined with (25) and (26) this yields the desired result. □

7.2 Proof of Theorem 3.3

Using Definition 3.1 of a size-biased pick we find

$$\begin{aligned}
&\hat{P}(\tilde{X}_1 \in dx_1, \ldots, \tilde{X}_{M(t)-1} \in dx_{M(t)-1}) \\
&\quad = \hat{E}[\hat{P}(\tilde{X}_1 \in dx_1, \ldots, \tilde{X}_{M(t)-1} \in dx_{M(t)-1} | X_1, \ldots, X_{M(t)}, M(t))] \\
&\quad = \hat{E}\left[\sum_{j=1}^{M(t)} \frac{X_j}{S_{M(t)}} \prod_{i=1}^{M(t)-1} 1_{\{X_{\alpha_j(i)}\in dx_i\}}\right] \\
&\quad = \frac{u_2\bar{s}_2}{\lambda(1-e^{-\theta t^{d+1}})}\sum_{m=1}^{\infty} P(M(t)=m|\sigma_2>t) E\left[\sum_{j=1}^{m} X_j \prod_{i=1}^{m-1} 1_{\{X_{\alpha_j(i)}\in dx_i\}} \Big| \sigma_2 > t, M(t) = m\right],
\end{aligned}$$

where the final equality follows from the same sequence of arguments as used in the proof of Theorem 3.2. Next, we note that

$$\begin{aligned}E[X_j(t)|\sigma_2 > t, M(t) = m] =& \int_0^\infty xP(X_j(t) \in dx|M(t) = m, \sigma_2 > t) = \int_0^\infty x\, g_t(x)dx \\ =& \int_0^{\gamma_d c_d^d(s_1)t^d} \frac{x^{1/d}}{d\gamma_d^{1/d} c_d(s_1)\phi(t)t} \exp\left[\frac{-u_2\bar{s}_2 x^{\frac{d+1}{d}}}{(d+1)\gamma_d^{1/d} c_d(s_1)}\right] dx \\ =& \frac{1}{\phi(t)tu_2\bar{s}_2}\left[1 - \exp\left(-\frac{u_2\bar{s}_2\gamma_d c_d^d(s_1)t^{d+1}}{d+1}\right)\right],\end{aligned}$$

and

$$\sum_{j=1}^m E\left[X_j \prod_{i=1}^{m-1} 1_{\{X_{\alpha_j(i)} \in dx_i\}} \Big| \sigma_2 > t, M(t) = m\right] = \sum_{j=1}^m E[X_j|\sigma_2 > t, M(t) = m] \prod_{i=1}^{m-1} g_t(x_i).$$

Together with Lemma 7.2 the result follows.

7.3 Proof of Proposition 3.4

First, we use Bayes' rule to find

$$P(\mathcal{E}_\zeta(t_1, \ldots, t_m)|\sigma_2 = t) = \frac{P(\sigma_2 \in dt|\mathcal{E}_\zeta(t_1, \ldots, t_m))\, P(\mathcal{E}_\zeta(t_1, \ldots, t_m))}{P(\sigma_2 \in dt)}. \tag{27}$$

Since $P(\sigma_2 \in dt)$ is given in (24) and $P(\mathcal{E}_\zeta(t_1, \ldots, t_m)) = \lambda^m e^{-\lambda\zeta}$, it remains to calculate $P(\sigma_2 \in dt|\mathcal{E}_\zeta(t_1, \ldots, t_m))$. It is easy to see that

$$P(\sigma_2 > t|\mathcal{E}_\zeta(t_1, \ldots, t_m)) = \exp\left(-u_2\bar{s}_2 V_{\mathcal{E}_t}\right) q(\zeta, t), \tag{28}$$

where $q(\zeta, t)$ is the probability that a type-2 mutation arises in a clone that is born in the interval (ζ, t). We find

$$\begin{aligned}q(\zeta, t) =& E\left[e^{-\theta\sum_{i=1}^{M(t-\zeta)}(t-T_i)^{d+1}}\right] \\ =& E\left[E\left[e^{-\theta\sum_{i=1}^{M(t-\zeta)}(t-T_i)^{d+1}} \Big| M(t-\zeta)\right]\right] \\ =& E\left[\phi(t-\zeta)^{M(t-\zeta)}\right] = e^{\lambda(t-\zeta)(\phi(t-\zeta)-1)},\end{aligned}$$

where the last expression is the generating function for the Poisson process. Together with (28) this yields now

$$\begin{aligned}P(\sigma_2 \in dt|\mathcal{E}_\zeta(t_1, \ldots, t_m)) =& -\frac{d}{dt}P(\sigma_2 > t|\mathcal{E}_\zeta(t_1, \ldots, t_m)) \\ =& e^{\lambda(t-\zeta)(\phi(t-\zeta)-1)} e^{-u_2\bar{s}_2 V_{\mathcal{E}_t}} \left[u_s\bar{s}_2 A_{\mathcal{E}_t} + \lambda\left(1 - e^{-\theta(t-\zeta)^{d+1}}\right)\right]\end{aligned}$$

Together with (24) and (18), we find now

$$P(\mathcal{E}_\zeta(t_1,\ldots,t_m)|\sigma_2=t)=\lambda^{m-1}\frac{e^{-\lambda[t\phi(t))-(t-\zeta)\phi(t-\zeta)]}}{\left(1-e^{-\theta t^{d+1}}\right)}e^{-u_2\bar{s}_2V_{\mathcal{E}_t}}\left[u_s\bar{s}_2A_{\mathcal{E}_t}+\lambda\left(1-e^{-\theta(t-\zeta)^{d+1}}\right)\right],$$

and hence performing the integration in

$$\hat{P}(M(\zeta)=m)=\int_{[0,\zeta]^m}\frac{1}{m!}P(\mathcal{E}_\zeta(t_1,\ldots,t_m)|\sigma_2=t)dt_1\ldots dt_m$$

yields the desired result.

7.4 Joint distribution of the process $\{M(r):0\le r\le t\}$

We present here the joint distribution of the process $\{M(r):0\le r\le t\}$, conditioned on $\sigma_2=t$, at multiple time points. Since the proof is similar to Proposition 3.4 we do not include it. For $0\le r\le r'\le t$ define

$$\hat{\phi}(t;r,r')=\int_r^{r'}e^{-\theta(t-y)^{d+1}}dy.$$

Then for any positive integer ℓ, sequence of time points $0<r_1\le\ldots\le r_\ell<t$ and non-negative integers $k_1\le k_2\le\ldots\le k_\ell$ we have that

$$\begin{aligned}&\hat{P}\left(M(r_1)=k_1,\ldots,M(r_\ell)=k_\ell\right)\\&=\left(\sum_{i=1}^{\ell}\frac{k_i-k_{i-1}}{\hat{\phi}(t;r_{i-1},r_i)}p_i+\lambda p_{\ell+1}\right)\frac{1}{\lambda}\prod_{j=1}^{\ell}\frac{\left(\lambda\hat{\phi}(t;r_{j-1},r_j)\right)^{k_j-k_{j-1}}}{(k_j-k_{j-1})!}e^{-\lambda\hat{\phi}(t;r_{j-1},r_j)},\end{aligned}$$

where for $1\le i\le\ell+1$,

$$p_i=\frac{e^{-\theta(t-r_i)^{d+1}}-e^{-\theta(t-r_{i-1})^{d+1}}}{1-e^{-\theta t^{d+1}}},$$

$r_0=0$, $k_0=0$, and $r_{\ell+1}=t$. Note that for each i, $0<p_i<1$ and $\sum_{i=1}^{\ell+1}p_i=1$, i.e. the p_i's form a probability vector. The above joint distribution is rather difficult to parse, so we describe how one would generate samples of the increments of the process. For $1\le i\le\ell$, set $X_i=M(r_i)-M(r_{i-1})$, then we can generate the values of the vector $X_1,\ldots,X_\ell$ under the measure $\hat{P}$ as follows. For each $1\le i\le\ell$ sample X_i according to a Poisson distribution with mean $\lambda\hat{\phi}(t;r_{i-1},r_i)$. Choose an integer I according to the probability vector $(p_1,\ldots,p_{\ell+1})$, if $I=i<\ell+1$ replace X_i with X_i+1. Note that in contrast to the setting of a Poisson process the random variables $X_1,\ldots,X_\ell$ are not independent under $\hat{P}$.

7.5 Proof of Corollary 4.1

$$\begin{aligned}\hat{P}(T_R^f > \tau) =& P\left(T_R^f > \tau | \sigma_2 = t\right) \\ =& \int_0^{c_d(s_1)t} P(T_R^f > \tau, R_l(\sigma_2) \in dr | \sigma_2 = t) dr \\ =& \int_0^{c_d(s_1)t} P(T_R^f > \tau | R_l(\sigma_2) \in dr, \sigma_2 = t) P(R_l(\sigma_2) \in dr | \sigma_2 = t) dr,\end{aligned}$$

where $R_l(t)$ is the radius of the local field surrounding the tumor at time t. The result follows from

$$P(T_R^f > \tau | R_l(\sigma_2) \in dr, \sigma_2 = t) = \exp\left(-\int_0^\tau \eta(r,s) ds\right) \tag{29}$$

and the conditional density of $R_l(\sigma_2)$ in (9).

7.6 Proof of Corollary 4.2

First, we note that

$$\begin{aligned}&P(T_R^p > \tau | M(t) = m, \sigma_2 = t) \\ &\quad = \int_{\mathbb{R}_+^{m-1}} P\left(T_R^p > \tau | \tilde{R}_1 \in dr_1, \ldots, \tilde{R}_{m-1} \in dr_{m-1}, M(t) = m, \sigma_2 = t\right) \cdots \\ &\qquad \cdots P\left(\tilde{R}_1 \in dr_1, \ldots, \tilde{R}_{m-1} \in dr_{m-1} | M(t) = m, \sigma_2 = t\right),\end{aligned} \tag{30}$$

where $\tilde{R}_i$ are the radii of the distant field clones, corresponding to their respective areas $\tilde{X}_i$ defined in Section 3.2. Recalling the definition of η in (12), we find

$$P\left(T_R^p > \tau | \tilde{R}_1 \in dr_1, \ldots, \tilde{R}_{m-1} \in dr_{m-1}, M(t) = m, \sigma_2 \in dt\right) = \exp\left(-\sum_{i=1}^{m-1} \int_0^\tau \eta\left(r_i, s\right) ds\right). \tag{31}$$

Recalling the Radon-Nikodym derivative $d\hat{P}/d\tilde{P}$ from Lemma 7.3, it is straight-forward to verify that

$$\frac{dP\left(t_1, \ldots t_m | M(t) = m, \sigma_2 = t\right)}{dP\left(t_1, \ldots t_m | M(t) = m, \sigma_2 > t\right)} = \frac{d\hat{P}}{d\tilde{P}} \frac{P(M(t) = m | \sigma_2 > t)}{P(M(t) = m | \sigma_2 = t)} = \frac{A_{\mathcal{E}_t} u_2 \bar{s}_2 t \phi(t)}{m(1 - e^{-\theta t^{d+1}})},$$

which allows us to derive the following expression (proceeding as in the proof of Corollary 3.3),

$$P\left(\tilde{X}_1 \in dx_1, \ldots, \tilde{X}_{m-1} \in dx_{m-1} | M(t) = m, \sigma_2 = t\right) = \prod_{i=1}^{m-1} g(x_i) dx_i.$$

Switching from the clone-areas $\tilde{X}_i$ back to the corresponding radii $\tilde{R}_i$, we find

$$P\left(\tilde{R}_1 \in dr_1, \ldots, \tilde{R}_{m-1} \in dr_{m-1} | M(t) = m, \sigma_2 = t\right) = (d\,\gamma_d)^{m-1} \prod_{i=1}^{m-1} r_i^{d-1} g(r_i^d \gamma_d) dr_i$$

From this, (31) and (30) we find

$$P\left(T_R^p > \tau | M(t) = m, \sigma_2 = t\right) = \left(d\,\gamma_d\,\Phi(\tau, t)\right)^{m-1}, \tag{32}$$

Finally, using Lemma 7.2,

$$\begin{aligned}\hat{P}(T_R^p > \tau) &= \sum_{m=1}^{\infty} P\left(T_R^p > \tau | M(t) = m, \sigma_2 = t\right) \hat{P}(M(t) = m) \\ &= \exp\left(-\lambda t \phi(t)\left(1 - d\gamma_d \Phi(\tau, t)\right)\right)\end{aligned}$$

References

[1] Danely P Slaughter, Harry W Southwick, and Walter Smejkal. field cancerization in oral stratified squamous epithelium. clinical implications of multicentric origin. *Cancer*, 6(5):963–968, 1953.

[2] Boudewijn JM Braakhuis, Maarten P Tabor, J Alain Kummer, C René Leemans, and Ruud H Brakenhoff. A genetic explanation of slaughter's concept of field cancerization evidence and clinical implications. *Cancer Research*, 63(8):1727–1730, 2003.

[3] Hong Chai and Robert E Brown. Field effect in cancer–an update. *Annals of Clinical & Laboratory Science*, 39(4):331–337, 2009.

[4] P. Armitage and R. Doll. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br. J. Cancer*, 11, 1957.

[5] G. Luebeck and S. Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *PNAS*, 99:15095–15100, 2002.

[6] N.L. Komarova, A. Sengupta, and M.A. Nowak. Mutation-selection networks of cancer initiation: Tumor suppressor genes and chromosone instability. *Journal of Theoretical Biology*, 223:433–450, 2003.

[7] F. Michor, Y. Iwasa, and M.A. Nowak. The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proc. Natl. Acad. Sci. USA*, 103:14931–14934, 2006.

[8] J. Schweinsberg. Waiting for n mutations. *Electronic Journal of Probability*, 13:1442–1478, 2008.

[9] Y. Iwasa, F. Michor, N. Komarova, and M. Nowak. Population genetics of tumor suppressor genes. *Journal of Theoretical Biology*, 233:15–23, 2005.

[10] D. Wodarz and N.L. Komarova. Can loss of apoptosis protect against cancer? *Trends Genet.*, 23:232–237, 2007.

[11] R. Durrett, D. Schmidt, and J. Schweinsberg. A waiting time problem arising from the study of multi-stage carcinogenesis. *Annals of Applied Probability*, 19:676–718, 2009.

[12] J. Foo, K. Leder, and F. Michor. Stochastic dynamics of cancer initiation. *Physical Biology*, 8:54–69, 2011.

[13] Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Martin A Nowak. Genetic progression and the waiting time to cancer. *PLoS Computational Biology*, 3(11):e225, 2007.

[14] N. Komarova. Spatial stochastic models for cancer initiation and progression. *Bull. Math. Biol.*, 68:1573–1599, 2006.

[15] M. Nowak, Y. Michor, and Y. Iwasa. The linear process of somatic evolution. *PNAS*, 100:14966–14969, 2003.

[16] T. Williams and R. Bjerknes. Stochastic model for abnormal clone spread through epithelial basal layer. *Nature*, 236:19–21, 1972.

[17] C. Thalhauser, J. Lowengrub, D. Stupack, and N. Komarova. Selection in spatial stochastic models of cancer: Migration as a key modulator of fitness. *Biology Direct*, 5:21, 2010.

[18] N. Komarova. Spatial stochastic models of cancer: Fitness, migration, invasion. *Mathematical Biosciences and Engineering*, 10:761–775, 2013.

[19] R. Durrett and S. Moseley. A spatial model for tumor growth. *Annals of Applied Probability, in press*, 2013.

[20] T. Liggett. *Stochastic interacting systems: contact, voter and exclusion processes.* Springer, 1999.

[21] M. Bramson and D. Griffeath. On the Williams-Bjerknes tumour growth model: I. *Annals of Probability*, 9:173–185, 1981.

[22] M. Bramson and D. Griffeath. On the Williams-Bjerknes tumor growth model: II. *Mathematical Proceedings of the Cambridge Philosophical Society*, 88:339–357, 1980.

[23] Erik A Martens and Oskar Hallatschek. Interfering waves of adaptation promote spatial mixing. *Genetics*, 189(3):1045–1060, 2011.

[24] Erik A Martens, Rumen Kostadinov, Carlo C Maley, and Oskar Hallatschek. Spatial structure increases the waiting time for cancer. *New journal of physics*, 13(11):115014, 2011.

[25] T. Antal, P. L. Krapivsky, and M. A. Nowak. Spatial evolution of tumors with successive driver mutations. *ArXiv e-prints*, 2013.

[26] R. Durrett, J. Foo, and K. Leder. Spatial Moran models II. Tumor growth and progression. *in revision*, 2013.

[27] R. Bertolusso and M. Kimmel. Modeling spatial effects in early carcinogenesis: Stochastic versus deterministic reaction-diffusion systems. *Math. Mod. Nat. Phenom.*, 7:245–260, 2012.

[28] R.A. Weinberg. *The Biology of Cancer [With DVD ROM]*. Taylor & Francis Group, 2013.

[29] A.M. Klein, D. P. Doupe, P. H. Jones, and B. D. Simons. Mechanism of murine epidermal maintenance: Cell division and the voter model. *Physical Review E*, 77(3), 2007.

[30] T.M. Liggett. *Interacting Particle Systems*. Classics in Mathematics Series. Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2005.

[31] R. Durrett. *Essentials of Stochastic Processes*. Springer Texts in Statistics. Springer, 2012.

[32] A. Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1:157–161, 2001.

[33] Camille Stephan-Otto Attolini and Franziska Michor. Evolutionary theory of cancer. *Annals of the New York Academy of Sciences*, 1168(1):23–51, 2009.

[34] Boudewijn JM Braakhuis, Maarten P Tabor, C René Leemans, Isaac van der Waal, Gordon B Snow, and Ruud H Brakenhoff. Second primary tumors and field cancerization in oral and oropharyngeal cancer: molecular techniques provide new insights and definitions. *Head & neck*, 24(2):198–206, 2002.

[35] C.R. Leemans, B.J.M. Braakhuis, and R.H. Brakenhoff. The molecular biology of head and neck cancer. *Nature Cancer Reviews*, 11:9–22, 2011.