Genetic Progression and the Waiting Time to Cancer

Niko Beerenwinkel[1¤*], Tibor Antal[1], David Dingli[1], Arne Traulsen[1], Kenneth W. Kinzler[2], Victor E. Velculescu[2], Bert Vogelstein[2,3], Martin A. Nowak[1]

1 Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts, United States of America, **2** Ludwig Center, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland, United States of America, **3** Howard Hughes Medical Institute, Johns Hopkins University, Baltimore, Maryland, United States of America

Cancer results from genetic alterations that disturb the normal cooperative behavior of cells. Recent high-throughput genomic studies of cancer cells have shown that the mutational landscape of cancer is complex and that individual cancers may evolve through mutations in as many as 20 different cancer-associated genes. We use data published by Sjöblom et al. (2006) to develop a new mathematical model for the somatic evolution of colorectal cancers. We employ the Wright-Fisher process for exploring the basic parameters of this evolutionary process and derive an analytical approximation for the expected waiting time to the cancer phenotype. Our results highlight the relative importance of selection over both the size of the cell population at risk and the mutation rate. The model predicts that the observed genetic diversity of cancer genomes can arise under a normal mutation rate if the average selective advantage per mutation is on the order of 1%. Increased mutation rates due to genetic instability would allow even smaller selective advantages during tumorigenesis. The complexity of cancer progression can be understood as the result of multiple sequential mutations, each of which has a relatively small but positive effect on net cell growth.

Citation: Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, et al. (2007) Genetic progression and the waiting time to cancer. PLoS Comput Biol 3(11): e225. doi:10.1371/journal.pcbi.0030225

Introduction

The current view of cancer is that tumorigenesis is due to the accumulation of mutations in oncogenes, tumor suppressor genes, and genetic instability genes [1]. Sequential mutations in these genes lead to most of the hallmarks of cancer [2]. Cancer research has benefited immensely from studies of uncommon inherited cancer syndromes that served to highlight the importance of individual genes in tumorigenesis [3]. Theoretical considerations have suggested that a handful of mutations, perhaps as few as three, may be sufficient for developing colorectal cancer [4,5]. This relatively small number is consistent with the standard model for colorectal tumorigenesis based on the identification of mutations in well-known cancer genes [6]. However, Sjöblom et al. [7] have recently determined the sequence of 13,000 genes in colorectal cancers and found that individual tumors contained an average of 62 nonsynonymous mutations. Extrapolating to the entire genome, it was estimated that individual colorectal cancers contain about 100 nonsynonymous mutations and that as many as 20 of the mutated genes in individual cancers might play a causal role in the neoplastic process [7].

Tumors arise from a process of replication, mutation, and selection through which a single cell acquires driver mutations which provide a fitness advantage by virtue of enhanced replication or resistance to apoptosis [8]. Each driver mutation thereby allows the mutant cell to go through a wave of clonal expansion. Along with drivers, passenger mutations, which do not confer any fitness advantage, are frequently observed. Passenger mutations arise in advantageous clones and become frequent by hitchhiking. The accumulation of ~100 mutations per cell is therefore the result of sequential waves of clonal expansion; the observed mutations mark the history of the cancer cell, including both drivers and passengers.

Genetic mutations can arise either due to errors during DNA replication or from exposure to genotoxic agents. The normal mutation rate due to replication errors is in the range of 10^{-10} to 10^{-9} per nucleotide per cell per division [9]. It is likely that the initial steps leading to cancer arise in cells with a normal mutation rate [10]. A normal mutation rate might also be sufficient to generate the large numbers of mutations in cancer given the many generations that the dominant cancer cell clone has gone through both before and after its initiating mutation [11–13]. However, it has also been argued that tumor cells have mutator phenotypes that accelerate the acquisition of mutations [14].

Mathematical modeling of carcinogenesis has had a rich history since its introduction more than 50 years ago [15–17]. The initial two-hit theory has evolved into more elaborate models incorporating multiple hits, rate-limiting events, and genomic instability [4,18–23]. Most models consider the stem cell at the base of the colonic crypt as the initial target for mutation, with the daughter cells giving rise to the adenoma

Editor: Greg Tucker-Kellogg, Lilly Singapore Centre for Drug Discovery, Singapore

Received July 3, 2007; **Accepted** October 2, 2007; **Published** November 9, 2007

Abbreviations: *APC*, adenomatous polyposis coli gene; *K-ras*, Kirsten rat sarcoma 2 viral oncogene homolog; *p53*, tumor protein 53

* To whom correspondence should be addressed. E-mail: niko.beerenwinkel@bsse.ethz.ch

¤ Current address: Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

Author Summary

Cancer is a disease of multicellular organisms that is characterized by a breakdown of cooperation between individual cells. The progression of cancer proceeds from a single genetically altered cell to billions of invasive cells through a series of clonal expansions. During tumorigenesis the cancer cells undergo replication and mutation, thereby increasing the size and invasiveness of the tumor. Recent sequencing projects of cancer cells suggest that mutations in up to 20 different genes might be responsible for driving an individual tumor's development. This insight contrasts with most mathematical models of cancer progression, which assume that the cancer phenotype is driven by mutations in only a few genes. We present a new mathematical model in which tumorigenesis is driven by mutations in many genes, most of which confer only a small selective advantage. Specifically, the progression of a benign tumor of the colon (adenoma) to a malignant tumor (carcinoma) is described by a Wright-Fisher process with growing population size. We explore the basic parameters of the model that are consistent with observed data. We also derive an analytical formula for the expected waiting time for the progression from benign to maligant tumor in terms of the population size, the mutation rate, the selective advantage, and the number of susceptible genes.

and progressively increasing the risk of malignant development [4,22].

The tumor data collected by Sjöblom et al. [7] show that the mutational patterns among colorectal cancers from different patients are diverse. This observation indicates that there may be many different mutational pathways that can lead to the same cancer phenotype. In the model described below, we assume that there are 100 potential driver genes and ask for the expected waiting time until one cell has acquired mutations in a given number, up to 20, of these genes. We assume that one or two initial mutations, perhaps together with losses or gains of large chromosomal regions [15,16], give rise to a benign tumor (adenoma) of $\sim$1 milligram or 10^6 cells (Figure 1). We model the progression of this adenoma to full-blown cancer over a period of five to 20 years [16], in which the adenoma grows to $\sim$1 gram, or 10^9 cells. Whether the whole population of cells is at risk for clonal expansion or whether a fraction of cells akin to stem cells drives growth of the adenoma is currently a subject of debate. This is important as cancer stem cells, as well as other factors such as geometric constraints on the architecture of the adenoma, may significantly reduce the effective population size and thereby impact the waiting time to cancer [24,25]. Note that it is not size that distinguishes a cancer from an adenoma; rather it is the ability of the cancer cells to invade through the underlying basement membrane and escape from its normal anatomical position.

We use the Wright-Fisher process [26] to model the somatic evolution of cancer in a colonic adenoma. We assume a cell turnover of one per day [27] and analyze the time to cancer as a function of the population size N, the per-gene mutation rate u, and the average selective advantage s per mutation. We present extensive simulation results as well as analytical approximations to the expected waiting time. The model offers a basic understanding of how the different evolutionary forces contribute to the progression of cancer.

Results

The mutation data are represented in a binary matrix of size 35×78, whose rows correspond to 35 tumor samples and whose columns correspond to the 78 candidate cancer genes identified by Sjöblom et al. [7] (Figure 2). A non-zero entry in cell (i, j) of this matrix indicates the presence of a mutation in gene j of tumor i. Tumors harbor between 1 and 20 mutated genes (mean = 6.5). Most of these genes ($66/78 = 85\%$) are mutated in at most three different tumors, resulting in highly diverse mutational patterns among the tumors. The notable exception are the three well-known cancer genes *APC*, *p53*, and *K-ras*, which were found mutated in 24, 17, and 16 tumors, respectively. We have analyzed partial correlations between genes, taking into account the small number of observations and multiple comparisons. Several pairs of genes were significantly correlated, most of them positively, but all correlations were weak and below 0.07 (Figure S1). From this data analysis, we conclude that in colon cancer, a very small number of genes are mutated in a large fraction of tumors. However, many other genes are involved in tumor progression, although each single gene is mutated only in a small subset of tumors without a clear pattern emerging.

For the purpose of mathematical modeling of tumorigenesis, we consider the presence of an adenoma. Adenoma formation probably requires the appearance of mutations in one or a few genes (in particular, *APC*) that are common to most tumors. We assume the occurrence of all subsequent mutations to be independent events. When any k out of $d = 100$ susceptible genes are mutated in a single cell, the cancer phenotype is considered to be attained. The first cells of this type mark the onset of an invasive tumor. The Wright-Fisher process is used to describe these evolutionary dynamics. Despite the large population size of up to $N = 10^9$ cells, we can efficiently compute estimates of the time to the first appearance of any k-fold mutant by simulation, because it

Figure 1. Schematic Representation of the Evolution of Cancer in a Colonic Adenoma

The adenoma grows from a population of 10^6 to 10^9 cells which accumulate mutations that drive phenotypic changes seen in cancer cells. Blue circles symbolize adenoma cells prior to accumulating the additional mutations that are the subject of modeling, green indicates cells that have acquired additional, but an insufficient number of mutations for malignancy, and red indicates cells with the number of mutations required for the cancer phenotype.

doi:10.1371/journal.pcbi.0030225.g001

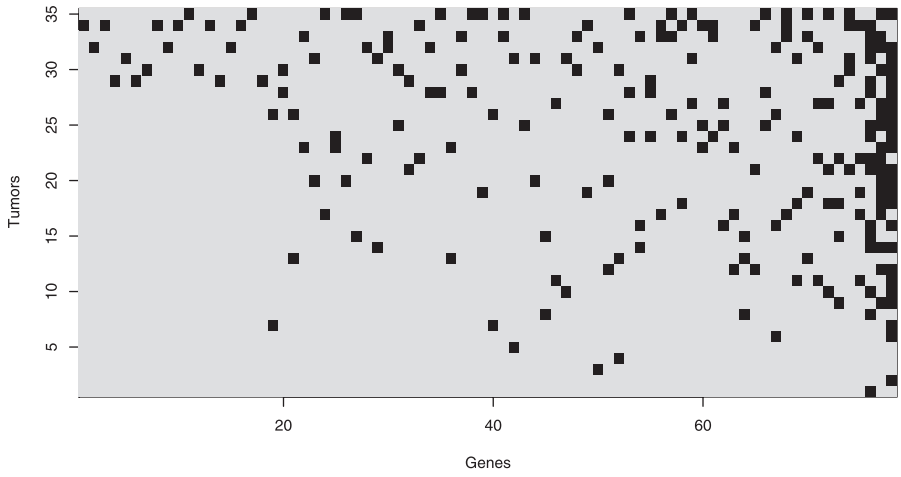

Figure 2. Mutational Patterns in 35 Late-Stage Colorectal Cancer Tumors from Sjöblom et al. (2006)

Matrix rows are indexed by tumors, columns are indexed by cancer-associated genes as identified by Sjöblom et al. (2006). Dark spots indicate mutated genes. Both tumors and genes have been sorted by an increasing number of mutations. The three genes mutated most often are *APC* (in 24 tumors; last column), *p53* (in 17 tumors; penultimate column), and *K-ras* (in 16 tumors; adjacent to *p53* column).

doi:10.1371/journal.pcbi.0030225.g002

suffices to trace the distribution of the $k+1$ mutant error classes in each generation. We assume a constant average selective advantage, s, for each mutation and a per-gene mutation rate, u. Figure 3 displays the typical behavior of this process in a single simulation. After a short initial phase in which the homogeneous wild-type population produces the first low-order mutants, a traveling wave is observed (Figure 3). Apparently, this distribution of error classes has constant variance and travels with constant velocity toward higher-order mutants. Thus, we expect the time until the first k-fold mutant appears to be linear in k. This conjecture is substantiated by simulations for a wide range of parameters (Figure S2) provided that mutations are advantageous ($s > 0$).

Within our model, the probability of developing cancer is equated with the probability of generating at least one k-fold mutant cell in the adenoma. For $k = 20$, this probability as a function of time is depicted in Figure 4. The expected time to the development of cancer increases with decreasing cell population size (hence the low risk of cancer associated with very small adenomas), with decreasing selective advantage, and with decreasing mutation rate. Thus, if the population at risk is a small subset composed of actively replicating stem cells, tumor progression will be slow. In contrast, an increased mutation rate due to genetic instability speeds up this process.

The simulations suggest that in a time frame of 5 to 15 years, cancer might develop in an adenoma of size 10^7 to 10^9 cells with a normal mutation rate of 10^{-7} per gene per cell division and a 1% selective advantage per mutation (Figure 4A). Alternatively, a higher mutation rate of 10^{-5} per gene per cell division would enable a smaller population of at-risk cells (10^5 to 10^7) and a smaller selective advantage (0.1%) to reach the required number of mutations in the same time interval (Figure 4B). However, for reasonable mutation rates, a completely neutral process ($s = 0$) predicts waiting times that are not consistent with the observed incidence of colon cancer, as would be expected (Figure 4, Figure S2).

Figure 5 generalizes these findings to different values of k by partitioning the parameter space of the model into regions of identical evolutionary outcomes. Each curve defines an instance of the Wright-Fisher process that results in a 10% chance of developing a k-fold mutant after 3,000 generations (or 8.2 years). These level curves define the parameter combinations that produce similar dynamics. For example, a small at-risk population is unlikely to generate a cancer requiring more than ten driver gene mutations unless the selective advantage for these mutations is large (see Discussion).

Based on the simulation results, we have derived an analytical approximation for the expected time to cancer. The key observation is that the distribution of error types follows a Gaussian (Figure 3). This approach leads to the expression

$$t_k = k\frac{\left(\log\frac{s}{ud}\right)^2}{s\log(N_{\text{init}}N_{\text{fin}})} \qquad (1)$$

for the expected waiting time, where k is the number of cancer-defining genes, d is the number of susceptible genes, u is the mutation rate, s the average selective advantage, and N_{init} and N_{fin} are the initial and final population sizes of the polyp, respectively (see Materials and Methods). The approximation is linear in k (Figure S2) and matches closely the observed behavior of the Wright-Fisher process, as long as $s > 0$ (Figures 3–5). The fit is analyzed quantitatively in Protocol S1. The expression for t_k highlights the strong effect of the selective advantage on tumorigenesis, and gives an explicit tradeoff between the evolutionary forces.

Discussion

Research over the past three decades has shown that cancer is an acquired genetic disorder [1]. The process of replication, mutation, and selection eventually leads to the appearance of tumors in multicellular organisms if they live long enough. Tumor cells accumulate many mutations in their evolutionary path [7,8,28], but not all mutations play a causal role in the evolution of the clone. If a gene is mutated in tumors

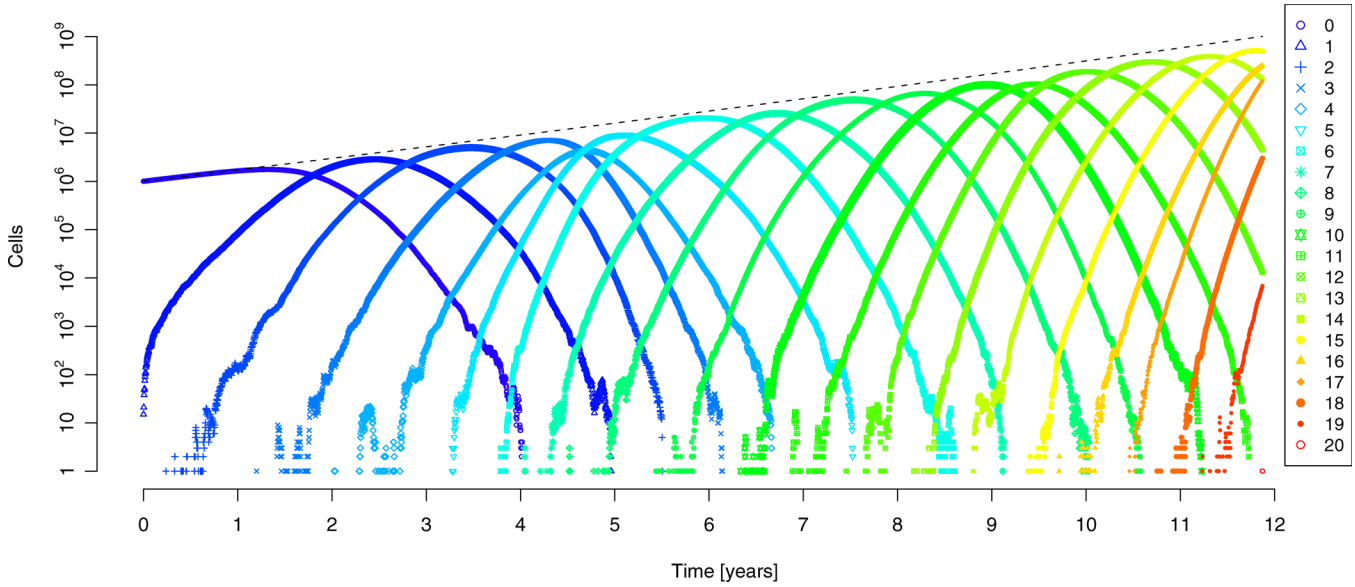

Figure 3. Evolution of Cancer Modeled by the Wright-Fisher Process

The distribution of cells in the error classes $N_0, \ldots, N_{20}$ is displayed in a single simulation over a time period of 12 years after which the first cell harboring 20 mutations appears. The total population size (dashed line) grows exponentially from 10^6 to 10^9 cells in this time period. Each cell has 100 susceptible genes, all of which are of wild-type initially. We further assumed a mutation rate of 10^{-7} per gene, a 1% selective advantage per mutation, and a turnover of 1 cell division per cell per day. Each error class has an approximately Gaussian distribution (after a short initial phase), but the introduction of each new mutant is subject to stochastic fluctuations.

doi:10.1371/journal.pcbi.0030225.g003

derived from different patients, it is less likely to be a passenger and more likely to provide the cell with a selective advantage, permitting it to expand and eventually dominate the population. Based on this reasoning, the data in Sjöblom et al. [7] suggest that as many as ~20 driver genes are mutated per tumor. The diverse mutational landscapes observed in tumor cells of the same tissue origin suggest that different mutations can have the same phenotypic effect. One plausible explanation for this observation is that genes are organized into intracellular pathways (signaling, metabolic, checkpoint, etc.), and the disturbance of these pathways drives tumorigenesis. Within each cell, every information transfer cascade requires functional proteins that are the products of distinct genes. Mutations in any one of the genes that code for proteins in a given pathway can complement each other and their genetic alterations can have similar phenotypic effects [1]. This view is supported by the observation that multiple hits in different genes of the same pathway in individual tumors are less frequent than expected [1].

In our model, we assume that each subsequent mutation has the same incremental effect on the fitness of the cell. In general, however, the impact of a specific mutation on the phenotype of the cell will depend on the genetic background. Gene interactions, or epistasis, can be positive or negative, and they can impose constraints on the order in which mutations accumulate [1]. In this case, the model parameter s may be regarded as the average fitness increase per mutation.

We have seen that a fitness increase of 1% per mutation may be enough for the Wright-Fisher model to generate dynamics that are consistent with the observed time scale of evolution from adenoma to carcinoma. However, genetic alterations associated with the initiation of colon cancer (such as those in *APC*) may have larger fitness advantages than those associated with tumor progression. The value or distribution of the fitness parameter s, which is unknown at present, will ultimately clarify the role of selection in this evolutionary process. We emphasize, however, that our estimates of s are based on the assumptions of the Wright-Fisher model. Thus, additional uncertainty in determining the role of selection is associated with violations of these assumptions by the biological system. Other models may lead to different estimates of the fitness associated with mutations, depending among other factors on the number of mutations necessary to reach the malignant phenotype and on the timescale [29].

In another simplifying abstraction, we have defined the tumor cell by the accumulation of $k = 20$ mutations in different driver genes. In reality, it is unlikely that any combination of 20 genes will induce the cancer phenotype. Our assumption is based on the observed cancer genotypes which fail to reveal a striking genetic signature of cancer cells. In this respect, our model provides lower bounds on the expected waiting time to cancer, as reaching a specific 20-fold mutant may take significantly longer.

These abstractions are important because all lesions begin with a small number of neoplastic cells. The simulations in Figure 5 show that cancers would never result from such small numbers of cells if 20 driver mutations were required and each mutation conferred only a small fitness advantage. It is likely that some of the early mutations (such as those in *K-ras*) increase fitness more than the average, allowing a small, initiating lesion to grow into an intermediate size lesion. Once a growth reaches this size, mutations with small fitness advantages can accumulate and eventually convert the tumor into a cancer.

The large population size of 10^9 cells would suggest that a

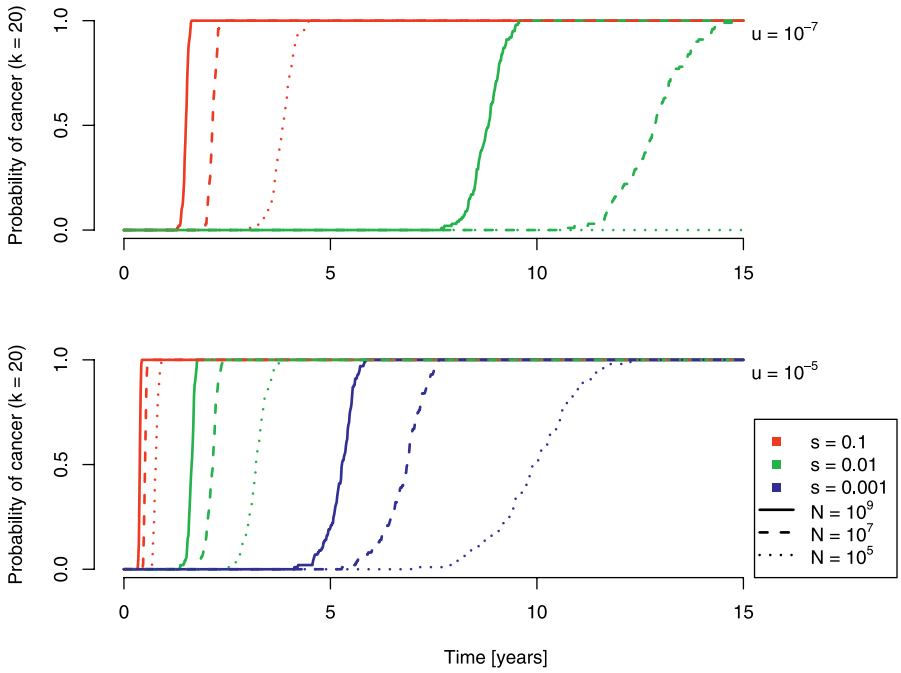

Figure 4. The Probability of Developing Cancer, Defined as the Occurrence of a Cell with Any 20 Mutated Genes Out of 100

Simulation results are displayed for three different population sizes (10^9, solid lines; 10^7, dashed lines; 10^5, dotted lines), three different selection coefficients (10%, red lines; 1%, green lines; 0.1%, blue lines), and two different mutation rates (10^{-7}, top; 10^{-5}, bottom).

doi:10.1371/journal.pcbi.0030225.g004

purely deterministic approximation to the Wright-Fisher process is reasonable. It turns out, however, that the stochasticity associated with generating mutants of each new type has a strong impact on the evolutionary dynamics (see Protocol S1). Therefore, a deterministic model of evolutionary dynamics will significantly underestimate the time to cancer. The closer approximation presented here exploits the regular behavior of the system of propagating a Gaussian distribution of error types and takes into account stochastic effects in determining the speed of this traveling wave. Thus, stochastic effects can play an important role even in very large populations.

Tumors derived from the same tissue exhibit considerable variability in their spectrum of mutations (Figure 2 and [7]). The number and type of mutations observed is the result of the size of the population at risk, the mutation rate, and the microenvironment of the evolving clone. The individual mutation rate can vary significantly due to genetic [27,30] and environmental effects (e.g., dietary fat intake, colonic bacterial flora, prior genotoxic therapy) [31,32]. These factors, expected to be different for every tumor, also contribute to the diversity of the mutational landscapes observed in tumors. It is also worth noting that the number of potential driver genes is likely to be an underestimate because the power of the Sjöblom et al. study to detect infrequent mutations was limited [7]. The study of larger numbers of tumors is likely to show that a few hundred different genes may function as drivers. This increase in potential drivers, however, will not have a substantial effect on the conclusions of the models derived here (Equation 1).

Most tissues in metazoans undergo turnover and are maintained by a population of tissue-specific stem cells that generally replicate at a slow rate and exhibit properties such as asymmetric division and immortal DNA strand co-segregation [33], perhaps to minimize the acquisition and retention of mutations. Although many tumors have cancer stem cells at their root [34] and colon cancer stem cells have been reported [24,25], it is an open question whether such cells arise solely due to the progressive accumulation of mutations in normal stem cells or because cells can re-acquire stem cell–like properties by mutation. The former scenario would suggest a much smaller effective population size, an important variable for modeling the evolution of cancer [4,22,35–37]. The colon has approximately 10^7 crypts, each one maintained by a small number of stem cells [27]. Initially, these stem cells constitute the overall population at risk, but the vast majority of patients with colon cancer develop tumors as the natural progression of mucosal adenomas [38]. Thus, adenoma formation can be regarded as a mechanism by which the population of cells at risk is increased and hence the probability of cancer in patients with multiple adenomas is dramatically increased. This is observed in familial adenomatous polyposis patients, who have inherited mutations of the *APC* gene.

Our model permits investigation of the impact of the relevant parameters of tumor evolution on a global scale. These parameters include the size of the population at risk, the mutation rate, and the fitness advantage conferred by specific mutations (Equation 1). The model suggests that the average waiting time for the appearance of the tumor is strongly affected by the fitness, s, conferred by the mutations, with the average waiting time decreasing roughly as $1/s$ (Figure S2). The mutation rate and the size of the population at risk contribute only logarithmically to the waiting time and hence have a weaker impact. Thus, the model of cancer progression presented here might add to the debate whether

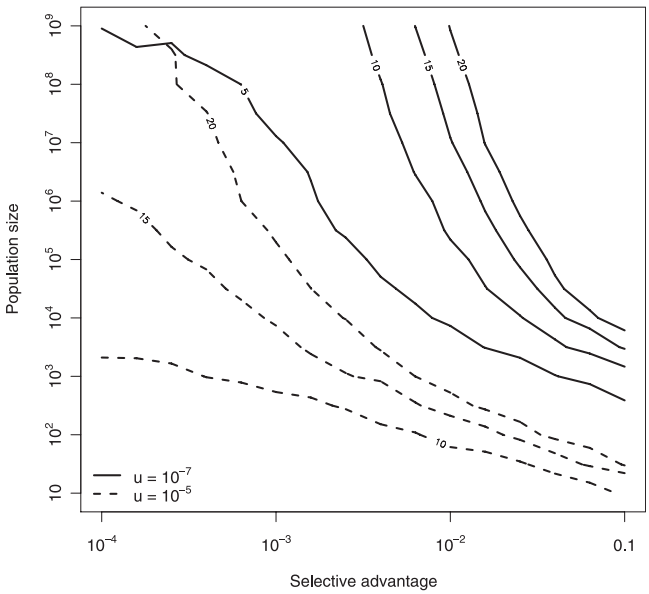

Figure 5. Level Curves of Identical Cancer Dynamics

Each curve connects points in parameter space (*x*-axis: selective advantage *s*, *y*-axis: population size *N*) with the same evolutionary outcome, namely a 10% chance of developing a *k*-fold mutant after 8.2 years (or 3,000 generations). The mutation rate is 10^{-7} (solid lines) and 10^{-5} (dashed lines), respectively. Curves are labeled with the number *k* of mutated genes that defines the cancer phenotype.

doi:10.1371/journal.pcbi.0030225.g005

selection [10,11] or mutation [39] is the dominant force in tumor development.

Finally, this model helps answer several questions about colorectal tumorigenesis that have long perplexed researchers and clinicians. Why is there so much heterogeneity in the times required for tumor progression among different patients? Why is there so much heterogeneity in the sizes and development times of tumors even within individual patients, such as those with familial adenomatous polyposis, if they all have the same initiating *APC* mutation? Why do cancers behave so differently with respect to their response to chemotherapeutic agents or radiation or their propensity to metastasize? Our model is compatible with the view that a few major mutational pathways, such as those involving *APC*, *K-ras*, and *p53*, endow relatively large increases in fitness that can allow tumors to grow to sizes compatible with further progression (Figure 5). However, the final course to malignancy will be determined by multiple mutations, each with a small and distinct fitness advantage, and these mutations occur stochastically. Every cancer will thereby be dependent on a unique complement of mutations that will determine its propensity to invade, its ability to metastasize, and its resistance to therapies. If this model is correct, then biological heterogeneity is a direct consequence of the tumorigenic process itself.

In our view, there is no reason to think that this model, or the data on which it was based, will be applicable only to colorectal cancers. Indeed, Sjöblom et al. [7] have identified similar mutational patterns in breast cancers, even though these tumors have completely different embryologic origins and are associated with distinct biological properties and predisposing factors. We therefore predict that the basic features of our model, i.e., a large number of potential drivers each of which contributes only a small fitness advantage, will apply to the progression of most common solid tumors. These tumors include those of the stomach, pancreas, bladder, lung, prostate, and kidney. It is unlikely that the model will apply to tumors that appear to have shorter waiting times, such as leukemias and lymphomas.

After completion and submission of the manuscript, we have learnt about related work recently published or being published [40–42]. These independent papers, which build on previous work published in [43], discuss a closely related mathematical model. In contrast to our work, these excellent contributions do not consider applications to the somatic evolution of cancer. Furthermore, we arrive at similar conclusions regarding the expected waiting time with a much more concise method than used in the other papers. We are grateful to Eric Brunet for bringing these references to our attention.

Methods

Data. The collection of tumor data has been described in [7]. Briefly, ~13,000 genes were sequenced from cancers of 11 patients with advanced colorectal cancers. Any mutant gene detected in this study was analyzed in an additional 24 patients with advanced cancers. Tumors with mismatch repair (MMR) deficiency were not included in this cohort, as MMR is known to increase the mutation rate by orders of magnitude and would complicate the analysis of mutations. Mutations were found in 519 genes, and, of these, 105 genes were found to be mutated in at least two independent tumors.

Statistical analysis. To test for dependencies between mutated genes, we calculated all 3,003 pairwise partial correlations between the 78 genes that were considered candidate drivers. Because the number of observed tumors is much smaller than the number of genes, we used the shrinkage method introduced in [44] for estimation.

Wright-Fisher process. We initially consider a colonic adenoma composed of 10^6 cells (~1 mm^3) that is growing exponentially to reach a size of 10^9 cells (~1 cm^3). Serial radiological observations show that the growth of unresected colonic adenomas is well-approximated by an exponential function [45]. The average growth rate determined in [45] implies that it takes ~11 years for an adenoma to grow from 10^6 to 10^9 cells. We consider an evolving cell population of size $N(t)$ in generation t. Population growth is modeled by assuming that growth is proportional to the average fitness $\langle w \rangle$ of the population, $N(t+1) - N(t) = \alpha \langle w \rangle N(t)$, where α is a constant ensuring the experimentally observed growth dynamics, and $N(0) = 10^6$. Although $\langle w \rangle$ changes slightly over time, the growth kinetics is still approximately exponential.

Each cell is represented by its genotype, which is a binary string of length $d = 100$ corresponding to the 100 potential driver genes. The population is initially homogeneous and composed of wild-type cells which are represented by the all-zeros string. In each generation, $N(t)$ genotypes are sampled with replacement from the previous generation. For large population sizes of 10^9 cells, it is not feasible to track the fate of each of the possible 2^{100} mutants in computer simulations. However, we are interested in the first appearance of any k-fold mutant in the system ($k = 20$). Thus, it suffices to trace the $k+1$ mutant error classes, i.e., the number of j-fold mutants $N_j(t)$ for each $j = 0, \ldots, k$, in each generation. With every additional mutation, we associate a selective advantage s. Thus, the relative fitness of a j-fold mutant is $w_j = (1+s)^j / \sum_{\ell=0}^{k} (1+s)^\ell x_\ell$, where $x_i = N_i / N$, and the average population fitness is $\langle w \rangle = \sum_{j=0}^{k} x_j w_j$. Ignoring back mutation, the probability of sampling a j-fold mutant is

$$\theta_j = \sum_{i=0}^{j} \binom{d-i}{j-i} u^{j-i} (1-u)^{d-j} w_i x_i(t),$$

where u is the mutation rate per gene. In each generation, the population is updated by sampling from the multinomial distribution

$$[N_0(t+1), ..., N_k(t+1)] \sim \frac{N(t)!}{N_0(t)! \cdot ... \cdot N_k(t)!} \prod_{j=0}^{k} \theta_j^{N_j(t)},$$

where $N(t)$ follows the above growth kinetics.

We use the discrete Wright-Fisher process rather than the continuous Moran process [26], which might seem more natural for cancer progression, because the Wright-Fisher process allows for efficient computer simulations even for very large population sizes. Both models behave similarly for large population sizes [26].

Analytical approximation. The large cell population size might suggest that one could consider a replicator equation in the limit as $N \to \infty$. However, this approach yields a Poisson distribution for the time-dependent relative frequencies $x_j(t)$ with parameter $\lambda = ud(e^{st} - 1) / s$, implying that the variance of x increases over time, which contrasts with the simulation results (Figure 3). The reason for this discrepancy is that, in the replicator equation, higher-order mutants with high fitness are instantaneously generated. Thus, the time for their expansion is underestimated compared to the waiting time in the stochastic system. See Protocol S1 for further discussion of this phenomenon.

To account for the stochastic fluctuations in the accumulation of k mutations, we model this process by decoupling mutation and selection (see Protocol S1 for mathematical details). Briefly, we assume that j-fold mutants are generated at a constant rate with increasing j. The Gaussian describing the distribution of mutant error classes has mean vt, variance σ^2, and travels with velocity $v = s\sigma^2$ (Figure 3). To determine v, we consider an (initially) exponentially growing subpopulation of j-fold mutants and calculate the expected time until one $(j + 1)$-mutant is produced. This leads to $v = 2s\log N / \left(\log \frac{s}{ud}\right)^2$, and for constant population size N, we obtain the approximation $t_k \approx k\left(\log \frac{s}{ud}\right)^2 / 2s\log N$ for the expected time to the first appearance of any k-fold mutant. The same waiting time in a population growing exponentially from initial size $N_{\text{init}} = N(0)$ to final size $N_{\text{fin}} = N(t_k)$ is equal to that in a constant population with effective population size $N = \sqrt{N_{\text{init}} N_{\text{fin}}}$. Thus the speed of the mutant wave in the growing population can be approximated by the average of the values corresponding to the initial and final population sizes. This leads to $t_k \approx k\left(\log \frac{s}{ud}\right)^2 / s\log(N_{\text{init}} N_{\text{fin}})$ for the waiting time in a population growing from N_{init} to N_{fin}. We will often restrict our attention to constant population sizes because of the equivalent waiting time in a constant population with effective size equal to the geometric mean of the initial and final population sizes.

Supporting Information

Figure S1. Histogram of $3003 = \binom{78}{2}$ Partial Correlations between All 78 Cancer-Associated Genes

Correlation coefficients have been computed from the 0/1 matrix displayed in Figure 2.

Found at doi:10.1371/journal.pcbi.0030225.sg001 (5 KB PDF).

Figure S2. Time T_k until, in 10% of Patients, k Genes Are Mutated

The waiting time T_k (y-axis) is plotted versus the number k of mutated genes (x-axis). Left panels correspond to a normal mutation rate of $u = 10^{-7}$, right panels to an increased mutation rate of $u = 10^{-5}$. Population sizes of 10^5 (top panels), 10^7 (middle panels), and 10^9 (bottom panels) are considered. The selective advantage per mutation varies among 0.1 (red lines), 0.01 (green), 0.001 (cyan), and 0 (purple).

Found at doi:10.1371/journal.pcbi.0030225.sg002 (11 KB PDF).

Protocol S1. PDF Document Entitled "Analytical Approximation for the Expected Waiting Time" Which Contains the Mathematical Details of the Model

Found at doi:10.1371/journal.pcbi.0030225.sd001 (152 KB PDF).

Acknowledgments

We are grateful to Tobias Sjöblom, Sian Jones, Jimmy Lin, Laura Wood, and Yoh Iwasa for helpful discussions.

Author contributions. KWK, VEV, and BV designed and performed experiments that formed the basis for the evaluation; NB, TA, DD, AT, and MAN developed the mathematical model; and NB, TA, DD, BV, and MAN wrote the paper.

Funding. Support from the US National Science Foundation/National Institutes of Health joint program in mathematical biology (NIH grant GM078986) is gratefully acknowledged. Genomics studies at Johns Hopkins are funded by the Virginia and D. K. Ludwig Fund for Cancer Research, the National Colorectal Cancer Research Alliance, The Maryland Tobacco Fund, and NIH grants CA43460, CA57345, CA62924, and CA121113. NB is funded by a grant from the Bill and Melinda Gates Foundation through the Grand Challenges in Global Health Initiative. The Program for Evolutionary Dynamics at Harvard University is sponsored by Jeffrey Epstein.

Competing interests. The authors have declared that no competing interests exist.

References

1. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. Nat Med 10: 789–799.
2. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.
3. Knudson AG (2002) Cancer genetics. Am J Med Genet 111: 96–102.
4. Luebeck EG, Moolgavkar SH (2002) Multistage carcinogenesis and the incidence of colorectal cancer. Proc Natl Acad Sci U S A 99: 15095–15100.
5. Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C (2003) The significance of unstable chromosomes in colorectal cancer. Nat Rev Cancer 3: 695–701.
6. Kinzler KW, Vogelstein B (1996) Lessons from hereditary colorectal cancer. Cell 87: 159–170.
7. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314: 268–274.
8. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158.
9. Kunkel TA, Bebenek K (2000) DNA replication fidelity. Annu Rev Biochem 69: 497–529.
10. Tomlinson I, Bodmer W (1999) Selection, the mutation rate and cancer: ensuring that the tail does not wag the dog. Nat Med 5: 11–12.
11. Wang T-L, Rago C, Silliman N, Ptak J, Markowitz S, et al. (2002) Prevalence of somatic alterations in the colorectal cancer cell genome. Proc Natl Acad Sci U S A 99: 3076–3080.
12. Tomlinson I, Sasieni P, Bodmer W (2002) How many mutations in a cancer? Am J Pathol 160: 755–758.
13. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. Nature 396: 643–649.
14. Loeb LA (2001) A mutator phenotype in cancer. Cancer Res 61: 3230–3239.
15. Nordling CO (1953) A new theory on cancer-inducing mechanism. Br J Cancer 7: 68–72.
16. Armitage P, Doll R (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 8: 1–12.
17. Armitage P, Doll R (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Br J Cancer 11: 161–169.
18. Moolgavkar SH, Knudson AG Jr (1981) Mutation and cancer: a model for human carcinogenesis. J Natl Cancer Inst 66: 1037–1052.
19. Nowak MA, Komarova NL, Sengupta A, Jallepalli PV, Shih Ie M, et al. (2002) The role of chromosomal instability in tumor initiation. Proc Natl Acad Sci U S A 99: 16226–16231.
20. Michor F, Iwasa Y, Nowak MA (2004) Dynamics of cancer progression. Nat Rev Cancer 4: 197–205.
21. Michor F, Iwasa Y, Lengauer C, Nowak MA (2005) Dynamics of colorectal cancer. Semin Cancer Biol 15: 484–493.
22. Little MP, Li G (2007) Stochastic modelling of colon cancer: is there a role for genomic instability? Carcinogenesis 28: 479–487.
23. Durrett R, Schmidt D, Schweinsberg J (2007) A waiting time problem arising from the study of multi-stage carcinogenesis. arXiv:0707.2057. Available: http://arxiv.org/abs/0707.2057. Accessed 11 October 2007.
24. O'Brien CA, Pollett A, Gallinger S, Dick JE (2007) A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature 445: 106–110.
25. Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, et al. (2007) Identification and expansion of human colon-cancer-initiating cells. Nature 445: 111–115.
26. Ewens WJ (2004) Mathematical population genetics. New York: Springer.
27. Potten CS (1998) Stem cells in gastrointestinal epithelium: numbers, characteristics and death. Philos Trans R Soc Lond B Biol Sci 353: 821–830.
28. Stoler DL, Chen N, Basik M, Kahlenberg MS, Rodriguez-Bigas MA, et al. (1999) The onset and extent of genomic instability in sporadic colorectal tumor progression. Proc Natl Acad Sci U S A 96: 15121–15126.
29. Dingli D, Traulsen A, Pacheco JM (2007) Stochastic dynamics of hematopoietic tumor stem cells. Cell Cycle 6: 461–466.
30. Friedberg EC (2003) DNA damage and repair. Nature 421: 436–440.
31. Slattery ML, Curtin K, Anderson K, Ma KN, Edwards S, et al. (2000) Associations between dietary intake and Ki-ras mutations in colon tumors: a population-based study. Cancer Res 60: 6935–6941.
32. Brink M, Weijenberg MP, De Goeij AF, Schouten LJ, Koedijk FD, et al.

(2004) Fat and K-ras mutations in sporadic colorectal cancer in The Netherlands Cohort Study. Carcinogenesis 25: 1619–1628.
33. Rambhatla L, Ram-Mohan S, Cheng JJ, Sherley JL (2005) Immortal DNA strand cosegregation requires p53/IMPDH-dependent asymmetric self-renewal associated with adult stem cells. Cancer Res 65: 3155–3161.
34. Reya T, Morrison SJ, Clarke MF, Weissman IL (2001) Stem cells, cancer, and cancer stem cells. Nature 414: 105–111.
35. van Leeuwen IM, Byrne HM, Jensen OE, King JR (2006) Crypt dynamics and colorectal cancer: advances in mathematical modelling. Cell Prolif 39: 157–181.
36. Michor F, Hughes TP, Iwasa Y, Branford S, Shah NP, et al. (2005) Dynamics of chronic myeloid leukaemia. Nature 435: 1267–1270.
37. d'Onofrio A, Tomlinson IP (2007) A nonlinear mathematical model of cell turnover, differentiation and tumorigenesis in the intestinal crypt. *J Theor Biol* 244: 367–374.
38. Winawer SJ (1999) Natural history of colorectal cancer. Am J Med 106: 3S-6S. Discussion 50S-51S.
39. Bielas JH, Loeb KR, Rubin BP, True LD, Loeb LA (2006) Human cancers express a mutator phenotype. Proc Natl Acad Sci U S A 103: 18238–18242.
40. Desai MM, Fisher DS (2007) Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176: 1759–1798.
41. Brunet E, Rouzine IM, Wilke CO (2007) The stochastic edge in adaptive evolution. arXiv:0707.3465. Available: http://arxiv.org/abs/0707.3465. Accessed 9 October 2007.
42. Rouzine IM, Brunet E, Wilke CO (2007) The traveling wave approach to asexual evolution: Muller's ratcher and speed of adaptation. arXiv:0707.3469. Available: http://arxiv.org/abs/0707.3469. Accessed 9 October 2007.
43. Rouzine IM, Wakeley J, Coffin JM (2003) The solitary wave of asexual evolution. Proc Natl Acad Sci U S A 100: 587–592.
44. Schafer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4: article32.
45. Welin S, Youker J, Spratt JS Jr (1963) The rates and patterns of growth of 375 tumors of the large intestine and rectum observed serially by double contrast enema study (Malmoe Technique). Am J Roentgenol Radium Ther Nucl Med 90: 673–687.

Analytical approximations for the expected waiting time

(Supporting information for N. Beerenwinkel, T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, M. A. Nowak, **Genetic progression and the waiting time to cancer**)

We derive analytical approximations for the expected waiting time for a cell with k mutations to appear. We consider the Wright-Fisher process for constant population size: we define the model in Section I, in Section II we present a simple argument which works only for weak selection, then in Section III we develop an approximation for strong selection. Finally in Section IV growing cell populations are investigated.

I. WRIGHT-FISHER PROCESS

Consider a population with a constant number N of cells. In every cell division mutations occur at rate u per locus. Each cell has d susceptible loci, and a mutation at each locus increases fitness by the same amount s. Thus, when j of the loci are mutated, the fitness of the cell is proportional to $(1+s)^j$. Let $N_j = N_j(t)$ be the number of cells with j mutations out of the d susceptible loci at time t, and $x_j = N_j/N$ be their relative frequency. We assume that the system evolves according to the Wright-Fisher model [1], where cells evolve in non-overlapping generations, and each cell independently chooses a parent cell from the previous generation with a probability proportional to the fitness of the parent. Each cell becomes identical to its parent apart from mutations which occur with probability u at each unmutated gene location. Consequently the probability of a configuration $[N_0(t+1), \ldots, N_d(t+1)]$ is given by the multinomial distribution

$$\frac{N!}{N_0(t)! \cdots N_d(t)!} \prod_{j=0}^{d} \theta_j^{N_j(t)} \tag{1}$$

with parameters

$$\theta_j = \sum_{i=0}^{j} \binom{d-i}{j-i} u^{j-i} (1-u)^{d-j} \frac{(1+s)^i x_i}{\sum_\ell (1+s)^\ell x_\ell}. \tag{2}$$

The parameter θ_j is the probability that a cell in the next generation will have j mutations. If the mutation rate is small $u \ll 1$ we can neglect multiple mutations, and θ_j simplifies to

$$\theta_j = \frac{(1+s)^j x_j}{\sum_\ell (1+s)^\ell x_\ell} + u(d-j+1) \frac{(1+s)^{j-1} x_{j-1}}{\sum_\ell (1+s)^\ell x_\ell}.$$

The first term is the probability to produce an additional cell of type j without mutation, while the second term is the probability that a cell of type $j-1$ mutates and produces a cell of type j. In the simulations we did not need to use this approximation.

II. DETERMINISTIC APPROACH

In the large N limit we may try to neglect stochastic fluctuations in order to obtain a deterministic equation [1]. We also assume that u and s are small, hence we only keep their leading order behavior. Considering $x_j(t)$ as a continuous variable in time we arrive at a system of ordinary differential equations

$$\dot{x}_j = u\left[(d-j+1)x_{j-1} - (d-j)x_j\right] + s x_j (j - \langle j \rangle) \tag{3}$$

where the dot represents the time derivative, and $\langle j \rangle$ is the average number of mutant loci at a given time,

$$\langle j \rangle = \sum_i i\, x_i(t). \tag{4}$$

The terms on the right hand side of (3) are easy to interpret. The first (gain) term describes cells with $j-1$ mutations becoming cells with j mutations by acquiring a new mutation at one of the $(d-j+1)$ possible loci. The second (loss) term similarly accounts for cells with j mutations undergoing a new mutation at one of the $(d-j)$ possible loci. The last term describes the effect of fitness, where each sub-population grows with a rate of their fitness advantage compared to the average fitness. Note also that densities remain normalized $\sum_j x_j = 1$ due to the $\langle j \rangle$ term.

We are interested in the time until the first cell with k mutations appears, i.e., until $x_k = 1/N$. If $k \ll d$, the number of available mutations is approximately d, and we have

$$\dot{x}_j = ud(x_{j-1} - x_j) + s x_j (j - \langle j \rangle), \tag{5}$$

a somewhat simpler system of coupled first order differential equations. The full solution is the Poisson distribution with the time dependent parameter $\lambda = \lambda(t)$,

$$x_j = \frac{\lambda^j e^{-\lambda}}{j!}, \quad \lambda = \frac{ud}{s}(e^{st} - 1). \tag{6}$$

This solution can be easily verified by substituting it back into (5). This solution describes a distribution with equal mean position and variance

$$\langle j \rangle = \operatorname{var} j = \lambda, \tag{7}$$

both growing exponentially in time for generic parameter values.

This behavior, however, is not supported by simulations, where we observe a traveling wave solution with constant speed and constant width (see Fig.1). The reason for the failure of this replicator description is the

following. The deterministic equation produces all types of mutants instantaneously, which then start to multiply, especially the ones with many mutations. This makes the distribution over j (or over t) much wider then in the simulations. In other words, N_0 is large enough for the deterministic equation to predict N_1 correctly, but then N_1 is relatively small when the first cell with $j = 2$ mutations arrives. Hence the fluctuations cannot be neglected, and the deterministic description fails to predict N_2 correctly.

Note, however, that without selection, *i.e.* for $s \to 0$ and $\lambda \to udt$, equation (6) becomes a good approximation. In this case, the time t_k to reach a k-fold mutant can be expressed from the condition $x_k(t_k) = 1/N$, as

$$t_k = \frac{-k}{ud}\, W\left[-\frac{k!^{1/k}}{kN^{1/k}}\right] \qquad \text{for } s \to 0, \tag{8}$$

where the Lambert W function is the inverse function of $f(x) = xe^x$ [2]. For example for $N = 10^9$ and $ud = 10^{-5}$ it gives $t_{20} \approx 3.5 \times 10^5$, while simulations result in $t_{20} \approx 5.6 \times 10^5$. For positive selection $s > 0$, however, we need to develop an alternative approximation, which we do in the next section.

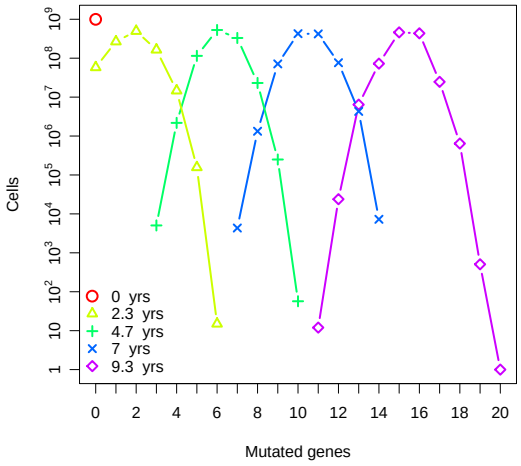

FIG. 1: Simulation results for the distribution of cells with given number of mutated genes at fixed times. The data can be very well approximated by a Gaussian wave traveling at constant speed to the right. The parameters of this simulation were $N = 10^9$, $s = 0.01$, $\mu = 10^{-7}$, $d = 100$, and generation time of one day.

III. WAVE-LIKE SOLUTION

Inspired by simulation results, we now develop a better approximation for the waiting time t_k. We decouple the evolution due to selection from the evolution due to mutation. We model the selection part as a deterministic process, but treat mutations stochastically.

First we consider only selection. For cell types already present in the system, we neglect the effect of mutation in the time evolution, since usually $s \gg ud$. Then the governing equation (3) simplifies to

$$\dot{x}_j = s x_j (j - \langle j \rangle), \tag{9}$$

where we extend the range of j to all integers. This equation has a Gaussian traveling wave solution

$$x_j = A \exp\left[-\frac{(j - vt)^2}{2\sigma^2}\right], \tag{10}$$

with constant speed v, and constant width σ. A continuously varying j would imply a normalization constant $A = 1/\sqrt{2\pi\sigma^2}$, and we use this value here as an approximation. Substituting solution (10) back into (9) yields a simple relationship between the speed and the width of the traveling wave of mutants,

$$v = s\sigma^2. \tag{11}$$

Now we have to consider the mutations which we have neglected so far. Notice that if we introduce each new type of mutant one after the other at a given speed, we also obtain (after some transient time) the solution (10) with the width given by (11). Simulations of the Wright-Fisher process support that $x_j(t)$ is a Gaussian (after an initial transient phase), that it has a constant width (see Fig. 1), and that the relationship (11) between the width and the speed holds.

Let us now derive an approximate expression for the speed v of the mutant wave in the stationary state. We need to know the average time τ at which the first new cell with $j + 1$ mutations appears after the birth of the first cell with j mutations. We assume that $\langle j \rangle$ does not change during this short time, and define the constant $\gamma = j - \langle j \rangle$. From (9) the density x_j initially grows exponentially in time [3],

$$x_j(t) = \frac{1}{N} e^{s\gamma t}, \tag{12}$$

where we also set $x_j(0) = 1/N$, as we start from a single mutant. We approximate the time τ as the time until, on average, one mutant is produced [4],

$$Nud \int_0^\tau x_j(t)dt = ud \int_0^\tau e^{s\gamma t} dt = \frac{ud}{s\gamma}(e^{s\gamma\tau} - 1) = 1\ ,$$

which leads to the speed of the mutant wave

$$v = \frac{1}{\tau} = \frac{s\gamma}{\log\left(1 + \frac{s\gamma}{ud}\right)} \approx \frac{s\gamma}{\log \frac{s\gamma}{ud}}\ . \tag{13}$$

As γ is typically of order one in our simulations, we assumed here that $s\gamma \gg ud$ is also true in the $s \gg ud$ limit.

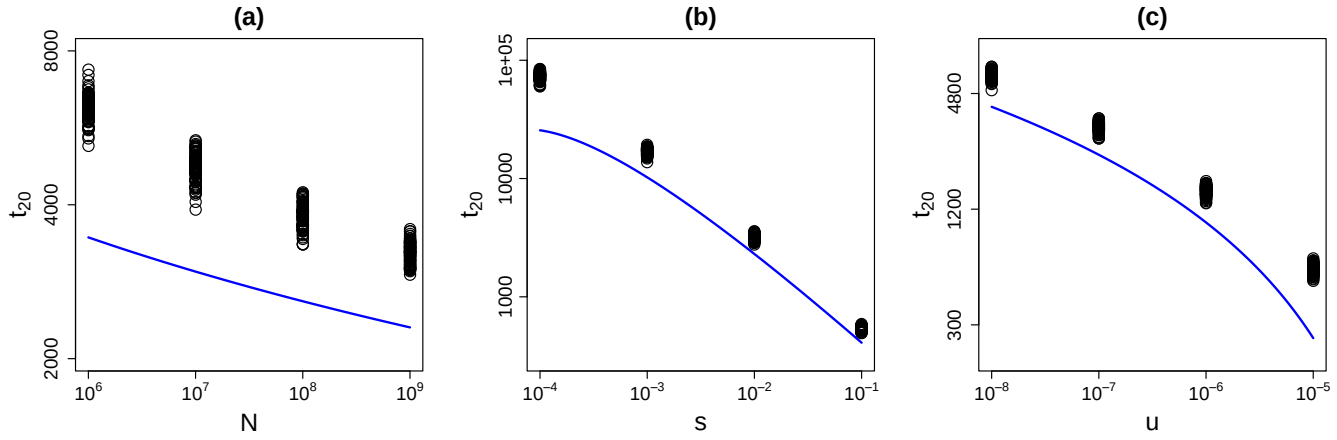

FIG. 2: Expected waiting time for a cell with 20 mutantions, t_{20}, as a function of (a) the population size N, (b) the selective advantage s per mutation, and (c) the per-locus mutation rate u. The circles are the results of 100 independent simulations at each parameter set. We always assumed $d = 100$ sensitive loci, and set $N = 10^9$ in (b) and (c), $s = 0.01$ in (a) and (c), and $u = 10^{-7}$ in (a) and (b). The solid curves correspond to the analytic approximation (17).

Next, we determine γ. Since $vt = \langle j \rangle$, at the moment when there is exactly one j cell, we have from (10) that

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\gamma^2}{2\sigma^2}\right) = \frac{1}{N} \tag{14}$$

and hence

$$\gamma = \sqrt{2}\sigma\sqrt{\log\frac{N}{\sqrt{2\pi\sigma^2}}} \approx \sqrt{2}\sigma\sqrt{\log N} = \sqrt{\frac{2v}{s}\log N}\ .$$

As σ is of order one, here we neglected $\log\sqrt{2\pi\sigma^2}$ next to $\log N$, and we also used (11) in the last step. Substituting γ into expression (13) for the speed we obtain

$$v = \frac{2s\log N}{\left[\log\left(\frac{s}{ud}\sqrt{\frac{2v}{s}\log N}\right)\right]^2}\ . \tag{15}$$

In the denominator we still have v inside the logarithm, which we approximate by the leading behavior $v \approx s$ to arrive at

$$v \approx \frac{2s\log N}{\left(\log\frac{s}{ud} + \frac{1}{2}\log\log N^2\right)^2} \approx \frac{2s\log N}{\left(\log\frac{s}{ud}\right)^2}\ , \tag{16}$$

where we also neglected the double logarithm term in the last step. This is our final formula for the speed of the wave. Using this expression for the speed we approximate the expected waiting time for the first k-fold mutant cell to appear as

$$t_k \approx \frac{k}{v} \approx k\,\frac{\left(\log\frac{s}{ud}\right)^2}{2s\log N} \tag{17}$$

In Figure 2, the dependence of t_k, for $k = 20$, on N, s, and u is analyzed by simulations of the Wright-Fisher model. The simple analytic argument given here leads to the appealing expression (17) for the expected waiting time, which is in good qualitative agreement with the simulation results for the Wright-Fisher process.

IV. GROWING POPULATION

Let us now study a population which grows exponentially from an initial size $N_{\rm init}$ to a final size $N_{\rm fin}$ during the evolution, that is $N(t) = N_{\rm init}e^{bt}$, where b is chosen such that $N(t_k) = N_{\rm fin}$. For the relative frequencies x_j, equation (10) is still valid, but the speed of the wave is no longer constant. Since the speed depends logarithmically on system size [see (16)], it grows linearly in time

$$v(t) = a\log N(t) = a(bt + \log N_{\rm init}) \tag{18}$$

where $a = 2s/[\log(s/ud)]^2$ is a constant. Hence the time at which the wave front reaches k mutations is given by

$$k = \int_0^{t_k} v(\tau)d\tau = at_k\,\frac{\log N_{\rm init} + \log N_{\rm fin}}{2} \tag{19}$$

which leads to

$$t_k \approx k\,\frac{\left(\log\frac{s}{ud}\right)^2}{s\log N_{\rm init}N_{\rm fin}} \tag{20}$$

for the waiting time for the k-fold mutant to appear. Note that this is also the waiting time in a constant population (17) with an effective population size $N_{\rm e} = \sqrt{N_{\rm init}N_{\rm fin}}$. Effective population sizes are frequently used in exponentially growing populations evolving according to the Wright-Fisher model [5].

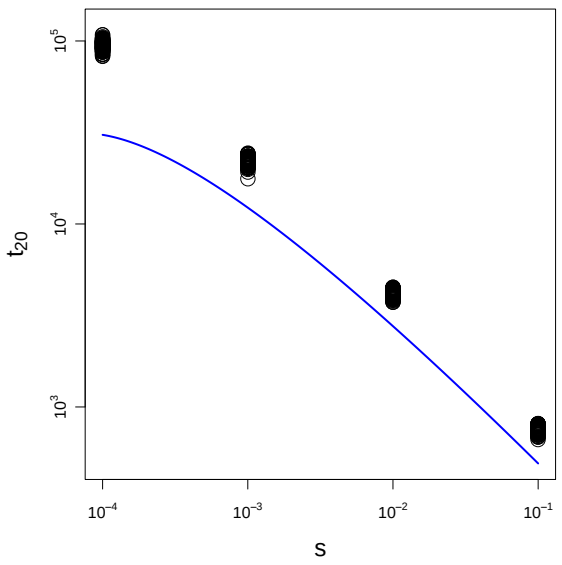

FIG. 3: Expected waiting time for a cell with $k = 20$ mutations, t_{20}, as a function of selection strength s, in a population which grows exponentially from size 10^6 to 10^9. The circles are simulation results of 100 runs for each s value, with mutation rate $u = 10^{-7}$ and $d = 100$. The solid curve is our analytical approximation (20).

In Figure 3 we compare the above formula to simulation results for a growing population. We conclude that our approximation works remarkably well also for growing populations.

[1] Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer, New York.

[2] Weisstein, Eric W. "Lambert W-Function." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/LambertW_Function.html

[3] Note that x_j eventually deviates from the early exponential growth and follows the Gaussian given by (10).

[4] More precisely we should take into account that a mutant survives only with a finite probability ρ, hence we should wait for $1/\rho$ mutants to appear in average. On the other hand the form assumed in (12) for the exponential growth is valid if we average over all possible trajectories, including mutants that go extinct. To obtain the average number of mutants under the condition that they survive we should multiply this expression also by $1/\rho$. Eventually, we have to multiply both sides of the condition (III) by $1/\rho$, which leaves the equation unchanged.

[5] Durrett, R. (2002) *Probability Models of DNA sequence evolution*, Springer-Verlag, New York.