Accumulation of driver and passenger mutations during tumor progression

Ivana Bozic[a,b], Tibor Antal[a,c], Hisashi Ohtsuki[d], Hannah Carter[e], Dewey Kim[e], Sining Chen[f], Rachel Karchin[e], Kenneth W. Kinzler[g], Bert Vogelstein[g,1], and Martin A. Nowak[a,b,h,1]

[a]Program for Evolutionary Dynamics, and [b]Department of Mathematics, Harvard University, Cambridge, MA 02138; [c]School of Mathematics, University of Edinburgh, Edinburgh EH9-3JZ, United Kingdom; [d]Department of Value and Decision Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan; [e]Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218; [f]Department of Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854; [g]Ludwig Center for Cancer Genetics and Therapeutics, and Howard Hudges Medical Institute at Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231; and [h]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Contributed by Bert Vogelstein, August 11, 2010 (sent for review May 26, 2010)

Major efforts to sequence cancer genomes are now occurring throughout the world. Though the emerging data from these studies are illuminating, their reconciliation with epidemiologic and clinical observations poses a major challenge. In the current study, we provide a mathematical model that begins to address this challenge. We model tumors as a discrete time branching process that starts with a single driver mutation and proceeds as each new driver mutation leads to a slightly increased rate of clonal expansion. Using the model, we observe tremendous variation in the rate of tumor development—providing an understanding of the heterogeneity in tumor sizes and development times that have been observed by epidemiologists and clinicians. Furthermore, the model provides a simple formula for the number of driver mutations as a function of the total number of mutations in the tumor. Finally, when applied to recent experimental data, the model allows us to calculate the actual selective advantage provided by typical somatic mutations in human tumors in situ. This selective advantage is surprisingly small—0.004 ± 0.0004—and has major implications for experimental cancer research.

genetics | mathematical biology

It is now well accepted that virtually all cancers result from the accumulated mutations in genes that increase the fitness of a tumor cell over that of the cells that surround it (1, 2). "Fitness" is defined as the net replication rate, i.e., the difference between the rate of cell birth and cell death. As a result of advances in technology and bioinformatics, it has recently become possible to determine the entire compendium of mutant genes in a tumor (3–9). Studies to date have revealed a complex genome, with ~40–80 amino acid changing mutations present in a typical solid tumor (6–10). For low-frequency mutations, it is difficult to distinguish "driver mutations"—defined as those that confer a selective growth advantage to the cell—from "passenger mutations" (11–13). Passenger mutations are defined as those which do not alter fitness but occurred in a cell that coincidentally or subsequently acquired a driver mutation, and are therefore found in every cell with that driver mutation. It is believed that only a small fraction of the total mutations in a tumor are driver mutations, but new, quantitative models are clearly needed to help interpret the significance of the mutational data and to put them into the perspective of modern clinical and experimental cancer research.

In most previous models of tumor evolution, mutations accumulate in cell populations of constant size (14–16) or of variable size, but the models take into account only one or two mutations (17–21). Such models typically address certain (important) aspects of cancer evolution, but not the whole process. Indeed, we now know that most solid tumors are the consequence of many sequential mutations, not just two. These tumors typically contain 40–100 coding gene alterations, including 5–15 driver mutations (6–9). The exploration of models with multiple mutations in growing tumor cell populations is therefore an essential line of investigation which has just recently been initiated (22, 23). In the model presented in this paper, we assume that each new driver mutation leads to a slightly faster tumor growth rate. This model is as simple as possible, because the analytical results depend on only three parameters: the average driver mutation rate u, the average selective advantage associated with driver mutations s, and the average cell division time T.

Tumors are initiated by the first genetic alteration that provides a relative fitness advantage. In the case of many leukemias, this would represent the first alteration of an oncogene, such as a translocation between *BCR* (breakpoint cluster region gene) and *ABL* (V-abl Abelson murine leukemia viral oncogene homolog 1 gene). In the case of solid tumors, the mutation that initiated the process might actually be the second "hit" in a tumor suppressor gene—the first hit affects one allele, without causing a growth change, whereas the second hit, in the opposite allele, leaves the cell without any functional suppressor, in accord with the two-hit hypothesis (24). It is important to point out that we are modeling tumor progression, not initiation (14, 15), because progression is rate limiting for cancer mortality—it generally requires three or more decades for metastatic cancers to develop from initiated cells in humans.

Our first goal is to characterize the times at which successive driver mutations arise in a tumor of growing size. We have employed a discrete time branching process (25) for this purpose because it makes the numerical simulations feasible. In a discrete time process, all cell divisions are synchronized. We present analytic formulas for this discrete time branching process and analogous formulas for the continuous time case whenever possible (*SI Appendix*). At each time step, a cell can either divide or differentiate, senesce, or die. In the context of tumor expansion, there is no difference between differentiation, death, and senescence, because none of these processes will result in a greater number of tumor cells than present prior to that time step. We assume that driver mutations reduce the probability that the cell will take this second course, i.e., that it will differentiate, die, or senesce, henceforth grouped as "stagnate." A cell with k driver mutations therefore has a stagnation probability $d_k = \frac{1}{2}(1-s)^k$. The division probability is $b_k = 1 - d_k$. The parameter s characterizes the selective advantage provided by a driver mutation.

Author contributions: I.B., T.A., R.K., B.V., and M.A.N. designed research; I.B., T.A., H.O., H.C., D.K., and S.C. performed research; I.B., T.A., H.O., H.C., D.K., S.C., R.K., and M.A.N. contributed new reagents/analytic tools; I.B., T.A., R.K., K.W.K., B.V., and M.A.N. analyzed data; and I.B., T.A., R.K., K.W.K., B.V., and M.A.N. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 18241.

[1]To whom correspondence may be addressed. E-mail: bertvog@gmail.com or martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1010978107/-/DCSupplemental.

When a cell divides, one of the daughter cells can receive an additional driver mutation with probability u. The point mutation rate in tumors is estimated to be ~5×10^{-10} per base pair per cell division (26). We estimate that there are ~34,000 positions in the genome that could become driver mutations (see *Materials and Methods* and *SI Appendix*). As the rate of chromosome loss in tumors is much higher than the rate of point mutation (14), a single point mutation is rate limiting for inactivation of tumor suppressor genes (when a point mutation in a tumor suppressor gene occurs, the other copy of that gene will likely be lost relatively quickly; ref. 27). The driver mutation rate is therefore ~3.4×10^{-5} per cell division ($\approx 2 \times 34{,}000 \times 5 \times 10^{-10}$), because u is the probability that one of the daughter cells will have an additional mutation. Our theory can accommodate any realistic mutation rate and the major numerical results are only weakly affected by varying the mutation rate within a reasonable range.

Experimental evidence suggests that tumor cells divide about once every 3 d in glioblastoma multiforme (28) and once every 4 d in colorectal cancers (26). Incorporating these division times into the simulations provided by our model leads to the dramatic results presented in Fig. 1. Though the same parameter values —$u = 3.4 \times 10^{-5}$ and $s = 0.4\%$—were used for each simulation, there was enormous variation in the rates of disease progression. For example, in patient 1, the second driver mutation had only occurred after 20 y following tumor initiation and the size of the tumor remained small (micrograms, representing $<10^5$ cells). In contrast, in patient 6, the second driver mutation occurred after less than 5 y, and by 25 y the tumor would weigh hundreds of grams ($>10^{11}$ cells), with the most common cell type in the tumor having three driver mutations. Patients 2–5 had progression rates between these two extreme cases.

We can calculate the average time between the appearance of successful cell lineages (Fig. 2). Not all new mutants are successful, because stochastic fluctuations can lead to the extinction of a lineage. The lineage of a cell with k driver mutations survives only with a probability approximately $1 - d_k/b_k \approx 2sk$. Assuming that $u \ll ks \ll 1$, the average time between the first successful cell with k and the first successful cell with $k+1$ driver mutations is given by

$$\tau_k = \frac{T}{ks}\log\frac{2ks}{u}. \qquad [1]$$

The acquisition of subsequent driver mutations becomes faster and faster. Intuitively, this is a consequence of each subsequent mutant clone growing at a faster rate than the one before. For example, for $u = 10^{-5}$, $s = 10^{-2}$, and $T = 4$ d, it takes on average 8.3 y until the second driver mutation emerges, but only 4.5 more years until the third driver mutation emerges. The cumulative time to accumulate k mutations grows logarithmically with k.

In contrast to driver mutations, passenger mutations do not confer a fitness advantage, and they do not modify tumor growth

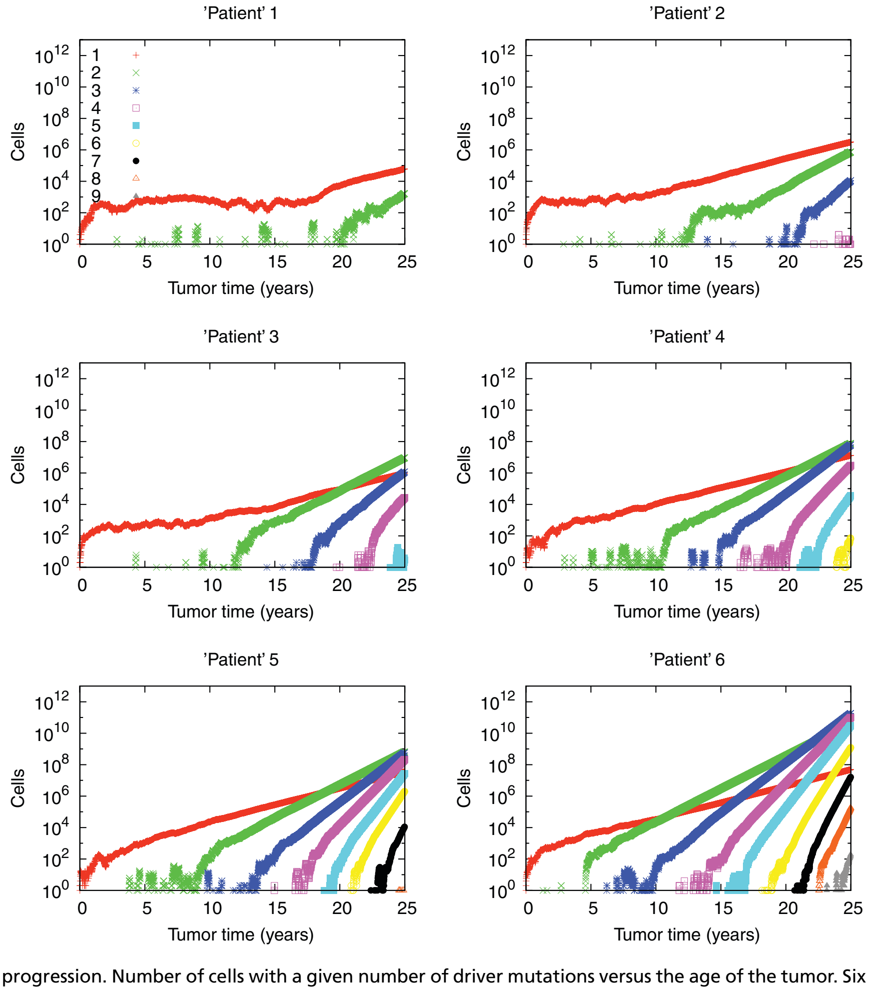

Fig. 1. Variability in tumor progression. Number of cells with a given number of driver mutations versus the age of the tumor. Six different realizations of the same stochastic process with the same parameter values are shown, corresponding to tumor growth in six patients. The process is initiated with a single surviving founder cell with one driver mutation. The times at which subsequent driver mutations arose varied widely among patients. After initial stochastic fluctuations, each new mutant lineage grew exponentially. The overall dynamics of tumor growth are greatly affected by the random time of the appearance of new mutants with surviving lineages. Parameter values: mutation rate $u = 3.4 \times 10^{-5}$, selective advantage $s = 0.4\%$, and generation time $T = 3$ d.

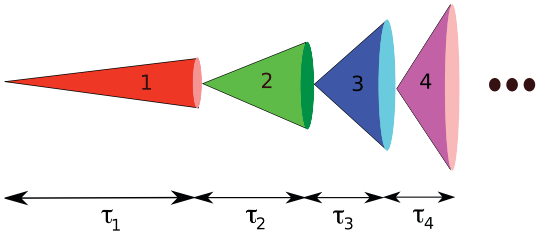

Fig. 2. Schematic representation of waves of clonal expansions. An illustration of a sequence of clonal expansions of cells with $k = 1$, 2, 3, or 4 driver mutations is shown. Here τ_1 is the average time it takes the lineage of the founder cell to produce a successful cell with two driver mutations. Similarly, τ_k is the average time between the appearance of cells with k and $k+1$ mutations. Eq. **1** gives a simple formula for these waiting times, which shows that subsequent driver mutations appear faster and faster. The cumulative time to have k driver mutations grows with the logarithm of k.

rates. We find that the average number of passenger mutations, $n(t)$, present in a tumor cell after t days is proportional to t, that is $n(t) = vt/T$, where v is the rate of acquisition of neutral mutations. In fact, v is the product of the point mutation rate per base pair and the number of base pairs analyzed. This simple relation has been used to analyze experimental results by providing estimates for relevant time scales (26).

Combining our results for driver and passenger mutations, we can derive a formula for the number of passengers that are expected in a tumor that has accumulated k driver mutations

$$n = \frac{v}{2s}\log\frac{4ks^2}{u^2}\log k. \qquad [2]$$

Here, n is the number of passengers that were present in the last cell that clonally expanded. Eq. **2** can be most easily applied to tumors in tissues in which there is not much cell division prior to tumor initiation. Otherwise, the expected number of passengers that accumulated in a precursor cell prior to tumor initiation would have to be included in the model, and this would be difficult to estimate.

We tested the validity of our model on two tumor types that have been extensively analyzed. Neither the astrocytic precursor cells that give rise to glioblastoma multiforme (GBM) (29) nor the pancreatic duct epithelial cells that give rise to pancreatic adenocarcinomas (30) divide much prior to tumor initiation (31, 32). Therefore, the data on both tumor types should be suitable for our analysis. Parsons et al. (8) sequenced 20,661 protein coding genes in a series of GBM tumors and found a total of 713 somatic mutations in the 14 samples that are depicted in Fig. 3. Similarly, Jones et al. (9) sequenced the same genes in a series of pancreatic adenocarcinomas, finding a total of 562 somatic mutations in the nine primary tumors graphed in Fig. 3. In both cases, we classified missense mutations as drivers if they scored high (false discovery rate ≤ 0.2) with the CHASM algorithm (33) and considered all nonsense mutations, out-of-frame insertions or deletions (INDELs), and splice-site changes as drivers because these generally lead to inactivation of the protein products (9). All other somatic mutations were considered to be passengers.

CHASM is a supervised statistical learning method that uses a Random Forest (34) to identify and prioritize somatic missense mutations most likely to that enhance tumor cell proliferation (drivers). The forest is trained on a positive class of ~2,500 missense mutations previously identified as playing a functional role in oncogenic transformation from the COSMIC database (35) and a negative class of ~4,000 random (passenger) missense mutations, which are synthetically generated with a computer algorithm. Mutations are represented by features derived from protein and nucleotide sequence databases, such as measures of evolutionary conservation, amino acid physiochemical properties, predicted protein structure, and annotations curated from the literature

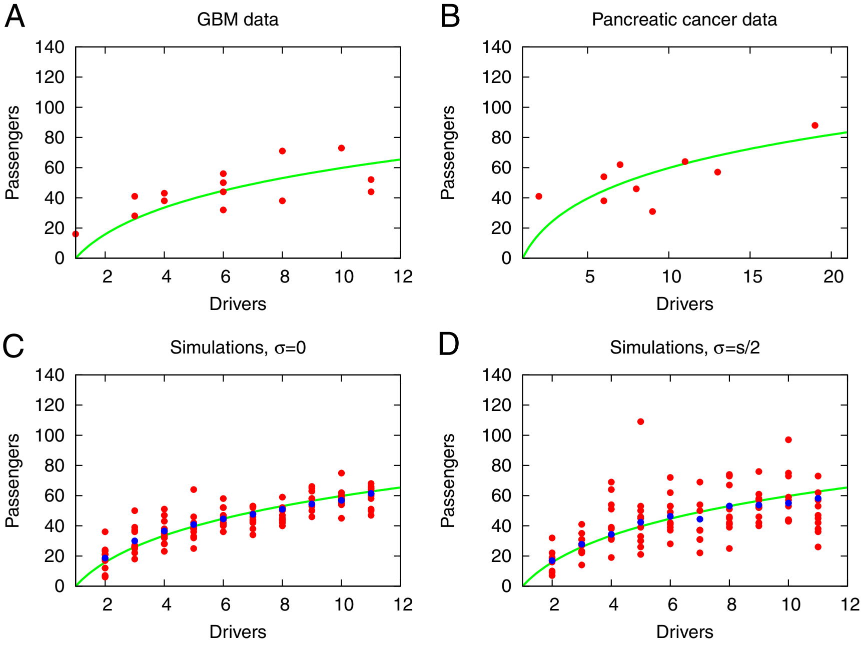

Fig. 3. Comparison of clinical mutation data and theory. Our theory provides an estimate for the number of passenger mutations in a tumor as a function of the number of driver mutations. Parameter values used in Eq. **2** and computer simulations were $s = 0.4\%$ and $u = 3.4 \times 10^{-5}$. (*A*) Eq. **2** (green line) fitted to GBM data. (*B*) Eq. **2** (green line) fitted to pancreatic cancer data. (*C*) Comparison of computer simulations and Eq. **2**. For each k between 2 and 10, the number of passengers that were brought along with the last driver in 10 tumors with k drivers is plotted. Blue circles represent averages from 100 simulations. (*D*) Comparison between computer simulations and Eq. **2** for selective advantage of the kth driver, s_k, taken from a Gaussian distribution with mean s and standard deviation $\sigma = s/2$. For each k between 2 and 10, the number of passengers that were brought along with the last driver in 10 tumors with k drivers is plotted. Blue circles represent averages from 100 simulations. Note that in *A*, the tumor with only one driver mutation has 16 passenger mutations, instead of the theoretically predicted zero. A possible reason for this discrepancy could be that the CHASM algorithm did not manage to classify all driver mutations as such, or perhaps that the ancestry of the founder cell of the tumor experienced a significant level of proliferation before the onset of neoplasia.

(from UniProtKB; ref. 36). There is nothing in the construction of the CHASM training set or features that mirrors the assumptions underlying the formulas derived here.

From Fig. 3 *A* and *B*, it is clear that the experimental results on both GBM and pancreatic cancers were in good accord with the predictions of Eq. **2**. A critical test of the model can be performed by comparison of the best-fit parameters governing each tumor type. It is expected that the average selective advantage of a driver mutation should be similar across all tumor types given that the pathways through which these mutations act overlap to a considerable degree. Setting the driver mutation rate to be $u = 3.4 \times 10^{-5}$, passenger mutation rate to be $v = 3.15 \times 10^7 \cdot 5 \times 10^{-10} \approx 0.016$, and fitting Eq. **2** to the GBM data using least squares analysis, we found that the optimum fit was given by $s = 0.004 \pm 0.0004$. Remarkably, using the same mutation rate in pancreatic cancers, we find that the best fit is given by a nearly identical $s = 0.0041 \pm 0.0004$. This consistency not only provides support for the model but also provides evidence that the average selective advantage of a driver is $s \approx 0.4\%$. For $u = 10^{-6}$ and $u = 10^{-4}$, we get $s \approx 0.65\%$ and $s \approx 0.32\%$, respectively. The fact that these estimates are not strongly dependent on the mutation rate supports the robustness of the model. Of course, we note that the reliability of the estimation of the passenger mutation rate v directly influences the reliability of estimating selection coefficients.

We conducted further testing of our model on data from two clinical studies (37, 38) of familial adenomatous polyposis (FAP) (39). FAP is caused by a germline mutation in one copy of the adenomatosis polyposis coli (*APC*) gene. Inactivation of the second copy of the *APC* gene in a colonic stem cell initiates the formation of a colonic adenoma. If untreated (by colectomy), patients with FAP develop adenomas while teenagers, but do not develop cancers until their fourth or fifth decades of life, by which time there are thousands of tumors per patient.

We performed computer simulations of the evolution of polyps in FAP patients. Assuming a constant number of susceptible stem cells and a constant rate of *APC* inactivation, new polyps in a patient are initiated at a constant rate. In simulations based on our model, we keep track of the number and size of all polyps in a patient and their change in time. We then compare simulation results with the clinical data from two studies (37, 38), focusing on three metrics of disease: (*i*) age distribution of FAP patients, (*ii*) number and size of visible polyps, and (*iii*) polyp growth rate.

To estimate the rate of polyp initiation in FAP, we estimate that there are ∼600 positions in the *APC* gene that, when mutated, could inactivate the *APC* gene product. However, the inactivation of *APC* in FAP patients more often happens by loss of heterozygosity (LOH) than by mutation—the ratio is ∼7 : 1 (for justification for these estimates, see *Materials and Methods*). Using the mutation rate per base pair per generation (26) of 5×10^{-10}, the rate of inactivation of *APC* is 2.4×10^{-6} per cell per generation. A typical human colon is ∼1.5 m long and has about 10^8 stem cells, each of which divides roughly once every week (40). In the clinical studies (37, 38), the authors only measure the number and size of polyps in the last 20 cm of the colon; the effective rate of *APC* inactivation in this part of the colon is ∼32 per stem cell generation, i.e., we estimate that 32 new polyps are initiated per week in this section of the colon. Note, however, that only a small fraction of these initiated cells will survive stochastic fluctuations.

The first study (37) included FAP patients that had at least five visible polyps, but no history of cancer. The number and size of their polyps was measured at baseline and a year later. To emulate the design of the study, each run of our simulation corresponded to one FAP "patient" <40 y old who had at least five visible polyps and no cancer (see *SI Appendix*). We then compared the age distribution of the patients in our simulation to the age distribution of patients in the study (37). Using the polyp initiation rate deduced above, mutation rate $u = 3.4 \times 10^{-5}$, generation time $T = 4$ d (26), and employing the selective advantage calculated from the GBM and pancreatic cancer data described above ($s = 0.004$), we find remarkable agreement between our model and the clinical data (Fig. 4). Our model predicted that patients would be entered into this study at an average of 25 y, with 35 polyps of average diameter 3.1 mm. The actual patients entered into the study had average age of 24 y, with 41 polyps of average diameter 3.2 mm. In comparison, if we keep mutation rate the same but emply a twofold lower or twofold higher value of s, then there is little agreement with the clinical data (e.g., age of diagnosis is either 38 or 14 y instead of the actual 24 y). We then used our model to predict the change in number and size of the polyps in these patients 1 y later. Our simulations predicted that the diameter and number of polyps would be 113% and 135% of the baseline values, respectively, whereas the diameter and number of polyps were 100% and 220% of baseline values in the actual patients.

We also modeled the results of a second study (38) that included 41 young FAP patients who had inherited alterations of the *APC* gene but had not yet developed polyps. These patients were followed for 4 y to determine when polyps first developed. Using the same simple assumptions noted above, our simulations predicted that 43% of these patients would develop at least one polyp within 4 y, and that the average diameter of polyps after 4 y would be 0.8 mm with standard deviation 0.9 mm. These predictions were remarkably similar to the data actually obtained, because 49% of the patients developed at least one polyp over the 4 y of observation and the average size of polyps was 0.9 mm with standard deviation 1.2 mm. However, our simulations underestimated the average number of polyps that developed (1.5 by the model, 6.7 in data), though there was a large variation in the number of polyps that developed in different patients (standard deviation of 12.5 polyps), complicating this metric.

Beerenwinkel et al. (22) previously modeled tumor evolution using a Wright–Fisher process. That model was specifically designed to model the evolution from a small adenoma to carcinoma, and it is not suitable for describing the dynamics of a population initiating with one or a small number of cells, as done

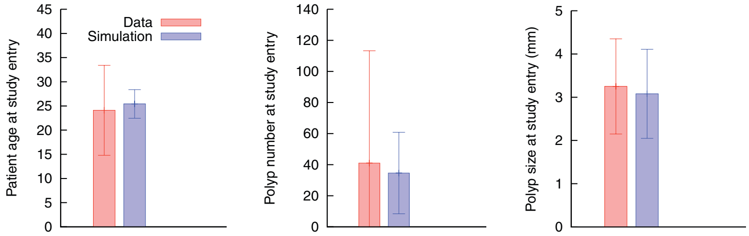

Fig. 4. Comparison of clinical FAP data and computer simulations of our model. A uniform random age <40 was picked first and only those patients who had at least five polyps and no history of cancer at the sampled age were retained. We compared the number and size of the polyps in these patients with the clinical data on number and size of polyps in FAP patients at study (37) entry. The age distribution of patients from the simulation was compared to the age distribution of patients in the study (37). Parameter values used in simulations are $s = 0.4\%$, $u = 3.4 \times 10^{-5}$, $T = 4$ days, and polyp initiation rate 32 per week. Error bars represent standard deviation.

here. Accordingly, the Beerenwinkel model does not address the long initial stages of the adenoma-carcinoma sequence (26) nor can it be used to model polyp development in FAP patients. Tumor progression in FAP patients has been previously modeled by Luebeck and coworkers (21, 41). At their rates, however, it takes a polyp about 60 y to grow to the average size of polyps reported in ref. 37. Our multistage model, where the growth rate is increasing with each new driver mutation, fits the observed polyp sizes well, providing strong and independent support for $s = 0.004$ as the selective growth advantage of a typical driver.

Like all models, ours incorporates limiting assumptions. However, many of these assumptions can be loosened without changing the key conclusions. For example, we assumed that the selective advantage of every driver was the same. We have tested whether our formulas still hold in a setting where the selective advantage of the kth driver is s_k, and s_ks are drawn from a Gaussian distribution with mean s and standard deviation $\sigma = s/2$. The simulations were still in excellent agreement with Eq. **2** (Fig. 3*D*). Similarly, we assumed that the time between cell divisions (generation time T) was constant. Nevertheless, Eq. **2**, which gives the relationship between drivers and passengers, is derived without any specification of time between cell divisions. Consequently, this formula is not affected by a possible change in T. Finally, there could be a finite carrying capacity for each mutant lineage. In other words, cells with one driver mutation may only grow up to a certain size, and the tumor may only grow further if it accumulates an extra mutation, allowing cells with two mutations to grow until they reach their carrying capacity and so on. It is reasonable to assume that the carrying capacities of each class would be much larger than $1/u$, which is approximately the number of cells with k mutations needed to produce a cell with $k + 1$ mutation. Thus, the times at which new mutations arise would not be much affected by this potential confounding factor.

Given the true complexity of cancer, our model is deliberately oversimplified. It is surprising that, despite this simplicity, the model captures several essential characteristics of tumor growth. Simple models have already been very successful in providing important insights into cancer. Notable examples include Armitage-Doll's multihit model (42), Knudson's two-hit hypothesis (24), and the carcinogenesis model of Moolgavkar and Knudson (43). The model described here represents an attempt to provide analytical insights into the relationship between drivers and passengers in tumor progression and will hopefully be similarly stimulating. One of the major conclusions, i.e., that the selective growth advantage afforded by the mutations that drive tumor progression is very small ($\sim 0.4\%$), has major implications for understanding tumor evolution. For example, it shows how difficult it will be to create valid in vitro models to test such mutations on tumor growth; such small selective growth advantages are nearly impossible to discern in cell culture over short time periods. And it explains why so many driver mutations are needed to form an advanced malignancy within the lifetime of an individual.

Materials and Methods

Oncogenes and Tumor Suppressor Genes Classifications. The COSMIC database contains sequencing information on 91,991 human tumors representing 353 different histopathologic subtypes (http://www.sanger.ac.uk/genetics/CGP/cosmic/). The database encompasses 105,084 intragenic mutations in 3,142 genes. Of these, 937 genes contained at least two nonsynynomous mutations, for a total of 97,567 mutations. We considered a gene to be a tumor suppressor if the ratio of inactivating mutations (stop codons due to nonsense mutations, splice-site alterations, or frameshifts due to deletions or insertions) to other mutations (missense and in-frame insertions or deletions) was >0.2. This criterion identified all well-studied tumor suppressor genes and classified 286 genes as tumor suppressors (*SI Appendix*). We considered a gene to be an oncogene if it was not classified as a tumor suppressor gene and either (*i*) the same amino acid was mutated in at least two independent tumors or (*ii*) >4 different mutations were identified (*SI Appendix*). This criterion classified 91 genes as oncogenes; the remaining 560 genes were considered to be passengers. There were an average of 13.6 different nucleotides mutated per oncogene.

Driver Positions in *APC*. In the entire *APC* gene, there are 8,529 bases encoding 2,843 codons. Of these bases, there are 3,135 bases representing 1,045 codons in which a base substitution resulting in a stop codon could occur. Only one-third of these 3,135 bases could mutate to a stop codon (e.g., an AAA could mutate to TAA to produce a stop codon, but a mutation to ATA would not produce a stop codon). Moreover, only one of the three possible substitutions at each base could result in a stop codon (e.g., a C could change to a T, A, G in general, but could only change to one of these bases to produce a stop codon). Therefore, the bases available for creating stop codons should be considered to be $3{,}135/9 = 348$ bases in the entire *APC* gene (i.e., 348 driver positions in *APC*). Insertions or deletions could also create stop codons in the *APC* gene. An estimate for the relative likelihood of developing an out-of-frame mutation can be obtained from our previous data (7–9). The number of nonsense mutations was 319, whereas the number of frameshift-INDELs was 235. Therefore, the total number of mutations leading to inactivating changes was 554, i.e., 174% of the number of nonsense codon-producing point mutations. The total number of driver positions in *APC* would therefore be 604 (174% of 348 nonsense driver positions).

Driver Positions in an Average Tumor Suppressor Gene. Assuming that the average tumor suppressor statistics follows that of the *APC*, and taking into account that the average number of base pairs in the coding region of the 23,000 genes listed in the Ensembl database (http://www.ensembl.org) is 1,604, we estimate that there are $604 \cdot 1{,}604/8{,}529 \sim 114$ driver positions in an average tumor suppressor gene.

Number of Driver Positions in the Genome. As noted above and in *SI Appendix*, we estimate that there are 286 tumor suppressor genes and 91 oncogenes in a human cell, and that on average each tumor suppressor gene can be inactivated by mutation at 114 positions and each oncogene can be activated in 14 positions. There are thus a total of 33,878 positions in the genome that could become driver mutations.

Relative Rate of LOH. The relative rate of LOH can be estimated from the data of Huang et al. (44). In this paper, mismatch repair (MMR)-deficient cancers were separated from MMR-proficient cancers. This separation is important because MMR-deficient cancers do not have chromosomal instability and they do not as often undergo LOH. We assume in all cases that the first hit was a somatic mutation of *APC*, and then the second hit could either have been LOH or mutation of a second allele. There were a total of 56 cancers analyzed in the study (44). Seven cancers had mutations in the other allele (i.e., two intragenic mutations), whereas the other 49 appeared to lose the second allele through an LOH event. Thus the relative rate of LOH vs. point mutation in *APC* is 7 : 1.

For further discussion and analysis of the model, see *SI Appendix*.

ACKNOWLEDGMENTS. This work is supported by The John Templeton Foundation, the National Science Foundation (NSF)/National Institutes of Health (NIH) (R01GM078986) joint program in mathematical biology, The Bill and Melinda Gates Foundation (Grand Challenges Grant 37874), NIH Grants CA 57345, CA 135877, and CA 62924, NSF Grant DBI 0845275, National Defense Science and Engineering Graduate Fellowship 32 CFR 168a, and J. Epstein.

1. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat Med* 10:789–799.
2. Greenman C, et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
3. Collins FS, Barker AD (2007) Mapping the cancer genome. *Sci Am* 296:50–57.
4. Ley TJ, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature* 456:66–72.
5. Mardis ER, et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 361:1058–66.
6. Sjoblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
7. Wood L, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
8. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812.
9. Jones S, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321:1801–1806.
10. Teschendorff AE, Caldas C (2009) The breast cancer somatic "muta-ome": Tackling the complexity. *Breast Cancer Res* 11:301.

11. Simpson AJ (2009) Sequence-based advances in the definition of cancer-associated gene mutations. *Curr Opin Oncol* 21:47–52.
12. Maley CC, et al. (2004) Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's Esophagus. *Cancer Res* 64:3414–3427.
13. Haber DA, Settleman J (2007) Cancer: Drivers and passengers. *Nature* 446:145–146.
14. Nowak MA, et al. (2002) The role of chromosomal instability in tumor initiation. *Proc Natl Acad Sci USA* 99:16226–16231.
15. Nowak MA, Michor F, Iwasa Y (2004) Evolutionary dynamics of tumor suppressor gene inactivation. *Proc Natl Acad Sci USA* 101:10635–10638.
16. Durrett R, Schmidt D, Schweinsberg J (2009) A waiting time problem arising from the study of multi-stage carcinogenesis. *Ann Appl Probab* 19:676–718.
17. Iwasa Y, Nowak MA, Michor F (2006) Evolution of resistance during clonal expansion. *Genetics* 172:2557–2566.
18. Haeno H, Iwasa Y, Michor F (2007) The evolution of two mutations during clonal expansion. *Genetics* 177:2209–2221.
19. Dewanji A, Luebeck EG, Moolgavkar SH (2005) A generalized Luria-Delbruck model. *Math Biosci* 197:140–152.
20. Komarova NL, Wu L, Baldi P (2007) The fixed-size Luria-Delbruck model with a nonzero death rate. *Math Biosci* 210:253–290.
21. Meza R, Jeon J, Moolgavkar SH, Luebeck G (2008) Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc Natl Acad Sci USA* 105:16284–16289.
22. Beerenwinkel N, et al. (2007) Genetic progression and the waiting time to cancer. *PLoS Comput Biol* 3:e225.
23. Durrett R, Moseley S (2010) The evolution of resistance and progression to disease during clonal expansion of cancer. *Theor Popul Biol* 77:42–48.
24. Knudson AG (1971) Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 68:820–823.
25. Athreya KB, Ney PE (1972) *Branching Processes* (Springer, New York).
26. Jones S, et al. (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci USA* 105:4283–4288.
27. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. *Nature* 396:643–649.
28. Hoshino T, Wilson CB (1979) Cell kinetic analyses of human malignant brain tumors (gliomas). *Cancer* 44:956–962.
29. Louis DN, et al. (2007) The 2007 WHO classification of tumors of the central nervous system. *Acta Neuropathol* 114:97–109.
30. Mimeault M, Brand RE, Sasson AA, Batra SK (2005) Recent advances on the molecular mechanisms involved in pancreatic cancer progression and therapies. *Pancreas* 31:301–316.
31. Kraus-Ruppert R, Laissue J, Odartchenko N (1973) Proliferation and turnover of glial cells in the forebrain of young adult mice as studied by repeated injections of ^{3}H-Thymidine over a prolonged period of time. *J Comp Neurol* 148:211–216.
32. Klein WM, Hruban RH, Klein-Szanto AJP, Wilentz RE (2002) Direct correlation between proliferative activity and displasia in pancreatic intraepithelial neoplasia (PanIN): Additional evidence for a recently proposed model of progression. *Mod Pathol* 15:441–447.
33. Carter H, et al. (2009) Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Res* 69:6660–6667.
34. Breiman L (2001) Random forest. *Mach Learn* 45:5–32.
35. Forbes SA, et al. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): A resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* 38(Database issue):D652–657.
36. UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* 38(Database issue):D142–148.
37. Giardiello FM, et al. (1993) Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *N Engl J Med* 328:1313–1316.
38. Giardiello FM, et al. (2002) Primary chemoprevention of familial adenomatous polyposis with sulindac. *N Engl J Med* 346:1054–1059.
39. Muto T, Bussey JR, Morson B (1975) The evolution of cancer of the colon and rectum. *Cancer* 36:2251–2270.
40. Potten CS, Booth C, Hargreaves D (2003) The small intestine as a model for evaluating adult tissue stem cell drug targets. *Cell Proliferat* 36:115–129.
41. Moolgavkar SH, Luebeck EG (1992) Multistage carcinogenesis: Population-based model for colon cancer. *J Natl Cancer Inst* 84:610–618.
42. Armitage P, Doll R (2004) The age distribution of cancer and a multi-stage theory of carcinogenesis. *Int J Epidemiol* 33:1174–1179.
43. Moolgavkar SH, Knudson AG (1981) Mutation and cancer: A model for human carcinogenesis. *J Natl Cancer Inst* 66:1037–1052.
44. Huang J, et al. (1996) APC mutations in colorectal tumors with mismatch-repair deficiency. *Proc Natl Acad Sci USA* 93:9049–9054.