Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation

Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation

Cristian Tomasetti[a,b,1], Bert Vogelstein[c,d,1], and Giovanni Parmigiani[a,b,1]

[a]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115; [b]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115; and [c]Ludwig Center for Cancer Genetics and Therapeutics and [d]Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231

Contributed by Bert Vogelstein, December 17, 2012 (sent for review October 3, 2012)

Although it has been hypothesized that some of the somatic mutations found in tumors may occur before tumor initiation, there is little experimental or conceptual data on this topic. To gain insights into this fundamental issue, we formulated a mathematical model for the evolution of somatic mutations in which all relevant phases of a tissue's history are considered. The model makes the prediction, validated by our empirical findings, that the number of somatic mutations in tumors of self-renewing tissues is positively correlated with the age of the patient at diagnosis. Importantly, our analysis indicates that half or more of the somatic mutations in certain tumors of self-renewing tissues occur before the onset of neoplasia. The model also provides a unique way to estimate the in vivo tissue-specific somatic mutation rates in normal tissues directly from the sequencing data of tumors. Our results have substantial implications for the interpretation of the large number of genome-wide cancer studies now being undertaken.

cancer evolution | mathematical modeling | stochastic processes | driver mutation | passenger mutation

The ever-growing amount of data originated by sequencing technologies has vastly enlarged our understanding of cancer genetics. A large number of somatic mutations are found in most solid tumors, and the great majority of these are "passengers," i.e., alterations that do not increase the selective growth advantage of the cells containing them, contrary to the so-called drivers (1). One fundamental question about these passenger mutations is their timing. A subset of these passenger mutations could in principle occur before the onset of neoplasia, defined as the occurrence of the first driver mutation (2). We have here modeled the process of accumulation of mutations and provide data suggesting that a substantial portion of the somatic mutations in typical adult human tumors arises before neoplastic development.

Fig. 1 shows the various phases of life during which somatic mutations can occur in a tissue's cell population that eventually develops a cancer. The shape of this process is "fish-like" as a result of the clonal bottlenecks that characterize each of these phases.

Development

A precursor cell, derived from the zygote, undergoes a clonal expansion from which a tissue is formed. In Fig. 1, this phase is represented by the head of the fish. Note that most of the mutations in this phase occur during embryonic or fetal life, as that is the time in which most clonal expansions leading to normal tissues occur.

Tissue self-renewal

Many healthy tissues regularly self-renew. These include those of the skin, gastrointestinal epithelium, hematopoietic system, and genitourinary tract. These renewals are represented by the body of the fish in Fig. 1, where vertical columns are used to depict each sequential renewal of the normal tissue. The average renewal time varies by cell type [about a week for the colon (3), possibly a month for hematopoietic stem cells (4)].

Tumorigenesis

A tumor is initiated by a driver mutation, i.e., a genetic alteration that increases the ratio of cell birth to cell death. In normal cell populations, even actively renewing ones, the long-term time average for this ratio should be 1. Once these initiated cells expand, successive clonal expansions occur with each new driver gene mutation. There is heterogeneity throughout this process, with clonal bottlenecks appearing as some clones predominate. Additional passenger mutations are accumulated with each clonal expansion. At any given point in tumor development, there will be at least some heterogeneity within the tumor as a result of anatomic constraints coupled with competing clone growth. This heterogeneity is depicted as the fish's tail.

Passenger mutations can occur at any time during these three phases. In Fig. 1, the brown-colored clones indicate the occurrence and possible expansion of cells with new passenger mutations. Even during the nonexpansionary self-renewal phase, genetic drift could produce the clonal expansion of a cell that has acquired passenger mutations. Such clones could later become extinct (shrinking back to zero in size). If the cell from which the cancer originates (represented in Fig. 1 by the left vertex of the cyan clone) were to originate from within a brown clone, then all tumor cells would contain the specific passenger mutations found in that brown clone.

Recent mathematical models have evaluated the accumulation of driver and passenger mutations during tumorigenesis (2, 5, 6). Although it has been hypothesized that some of the passenger mutations in tumors may occur before tumor initiation, the precancer phases have not been evaluated, or even modeled, in depth. As Fig. 1 indicates, the tail of the fish is only part of the tale.

To capture this aspect of somatic mutagenesis, we have formulated a mathematical model in which all relevant phases have been included. The model is based on widely accepted, straightforward assumptions. A model is useful only if it illuminates mechanisms and makes nonobvious, testable predictions that can guide future experimental research. Our model makes three predictions:

1. The number of somatic mutations in tumors of self-renewing tissues should be positively correlated with the age of the patient at diagnosis.
2. A large fraction of the somatic mutations in cancers of self-renewing tissues arises before tumor initiation.
3. It should be possible to estimate the background somatic mutation rate from the number of somatic mutations present in a tumor biopsy.

These predictions are tested as described below. Their confirmation leads to the conclusion that half or more of the

Author contributions: C.T. conceived the idea; C.T., B.V., and G.P. designed research; C.T. performed research and provided the mathematical modeling/analysis; C.T., B.V., and G.P. analyzed data; and C.T., B.V., and G.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221068110/-/DCSupplemental.

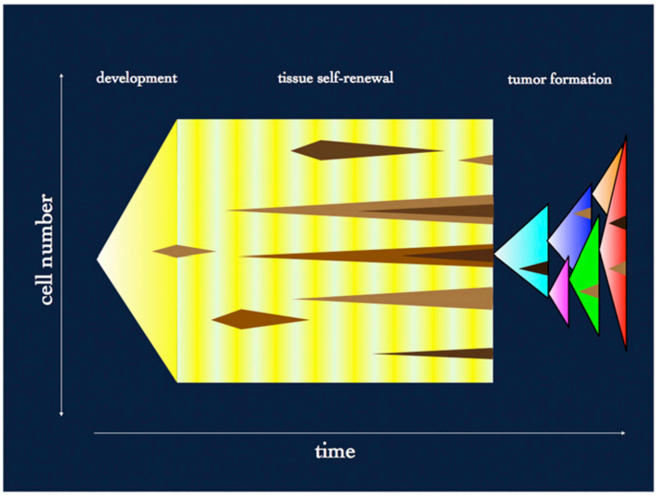

Fig. 1. The "Fish," a schematic representation of the different phases in which somatic mutations occur in a tissue giving rise to a cancer. Starting from a single precursor cell, a tissue is created via clonal expansion (head of the fish). The tissue is then subjected to periodic self-renewals (body of the fish). During development and tissue renewal, passenger mutations occur randomly, undergo clonal expansions (various brown clones), and either go extinct or expand as successive passenger mutations accumulate. A driver gene mutation can initiate a tumor cell clone, which then can expand through subsequent driver mutations, eventually yielding a clinically detectable tumor mass (fish's tail, where each clonal expansion driver by a new driver mutation is indicated by a different color). Passenger mutations occur during this phase as well.

somatic mutations in tumors of self-renewing tissues arise before tumor initiation.

Results

Number of somatic mutations found in cancer tissues correlates with age. As can be seen in Fig. 1, the tissue self-renewal phase should give rise to some fraction of the somatic mutations present in a tumor. The length of this renewal phase is directly proportional to the age of the patient when the first initiating driver mutation occurred. In self-renewing cell populations, the model thus predicts a positive correlation between the number of somatic mutations found in the tumor and the age of the patient at diagnosis (assuming the time from tumor initiation to age of diagnosis is relatively constant).

To test this prediction, we analyzed four large whole-exome sequencing datasets publicly available on The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) websites (Fig. 2): chronic lymphocytic leukemia (CLL) (109 patients), uterine corpus endometrioid carcinoma (229 patients), colorectal cancer (224 patients) (7), and pancreatic cancer (114 patients).

In each dataset, we removed tumors that were outliers, i.e., had very high numbers of mutations, as these were likely to be repair-deficient cancers with much higher rates of mutation than the other tumors (*Materials and Methods*). For CLL, there was a highly significant correlation between the number of mutations and age at diagnosis (Fig. 2*A*, $P = 0.0029$). Similarly, there were statistically significant positive correlations for uterine (Fig. 2*B*, $P = 0.0083$) and colorectal cancers (Fig. 2*C*, $P = 0.009$). Importantly, the tumor stage was not related to the age of the patient at diagnosis.

We also used a robust generalized linear model (*Materials and Methods*), where no patient (outlier) was removed, to test for the association between the number of mutations and age at diagnosis. Again, for CLL, uterine, and colorectal cancers, there was a highly significant association (Fig. 2*A*: $P = 7.45 \times 10^{-11}$; Fig. 2 *B* and *C*: $P < 2 \times 10^{-16}$).

Additional evidence supporting our prediction was provided by pancreatic cancers. It is known that normal pancreatic ductal epithelial cells, which are the precursors to pancreatic ductal adenocarcinomas, do not self-renew (8). In the graphical depiction of our model in Fig. 1, this would mean that there is no body to the fish; just the development (head) and tumor (tail) phases are present. Mathematically, our model thereby predicts that there should be no correlation between age at diagnosis and number of mutations in pancreatic ductal adenocarcinomas. Indeed, this prediction was verified in Fig. 2*D*: an approximately horizontal line in the plot of age vs. mutation number, with no significant correlation ($P = 0.18$), or association ($P = 0.38$), between age at diagnosis and mutation number. Further support for our prediction is provided by a small study of acute myeloid leukemia, and prior studies of pediatric tumors such as neuroblastoma and medulloblastoma, where a correlation between age and mutation number was noted (9–12).

Fraction of somatic mutations found in cancer tissues that originated before cancer initiation. The average number of somatic mutations in patients of various ages can be estimated by the regressions depicted in Fig. 2 (*Materials and Methods*). As shown in Table 1, it is substantially higher in 85-y-old patients than in 25-y-old patients: 24.16 vs. 10.09 (CLL), 96.48 vs. 45.96 (uterine), and 121.38 vs. 50.2 (colorectal), consistent with our model's prediction. As tumor stage was not related to the age of the patient at diagnosis, our analysis strongly supports the idea that a large portion of passenger mutations accumulates before the onset of neoplasia. Note that we do not need to distinguish passenger mutations from total mutations in our calculations, as it is widely accepted that the vast majority of somatic mutations are passengers (13).

What fraction of somatic mutations in a tumor actually arises in the precursor cells before tumor initiation? The number (and fraction) of mutations that occurred before tumor initiation can be estimated by subtracting the number of mutations that occurred during tumor progression (the tail of the fish) from the total number. To estimate this value, we need to know the average time it takes for a tumor to reach detection size. It has been estimated that colorectal cancer requires an average of 25 y (2), whereas leukemias take 7 y (14). For uterine cancer a value of 10 y is assumed (15). By using the regressions depicted in Fig. 2, we estimate the average number of somatic mutations present in a 7-y-old CLL, 10-y-old uterine, and 25-y-old colorectal cancer patient to be 5.86, 33.3, and 50.2, respectively. As shown in Table 2, our calculations suggest then that 68%, 57%, and 51% of the passenger somatic mutations in the median-age patient with CLL, uterine, or colorectal cancer, respectively, developed before tumor initiation. The median age at diagnosis was 61, 63, and 69 y in the CLL, uterine, and colorectal cancer datasets, respectively. Equivalent results are obtained if the regression slopes are used to determine the number of somatic mutation accumulated in a median-age patient, where the average number of years required for tumor progression has been subtracted. If uterine corpus endometrioid carcinoma took instead 20 y on average to reach detection, the average number of passenger mutations occurring during tumor progression would be 41.75, and therefore 46.5% of the passenger somatic mutations in the median-age patient would have developed before tumor initiation.

The result in ref. 9 further supports our prediction.

Estimating tissue-specific somatic mutation rates in vivo. Using our model and the slopes of the regressions in Fig. 2, we can also estimate the in vivo tissue-specific somatic point mutation rates.

The expected value for the number $X_S(t)$ of passenger mutations that originated in the precancer phase during tissue self-renewal (*Materials and Methods*) is estimated to be

$$E[X_S(t)] = S\,u\,t, \qquad [1]$$

where E is the expectation operator, S is the total number of DNA bases sequenced, u is the probability of a point mutation per base per cell division, and t is the number of times the tissue

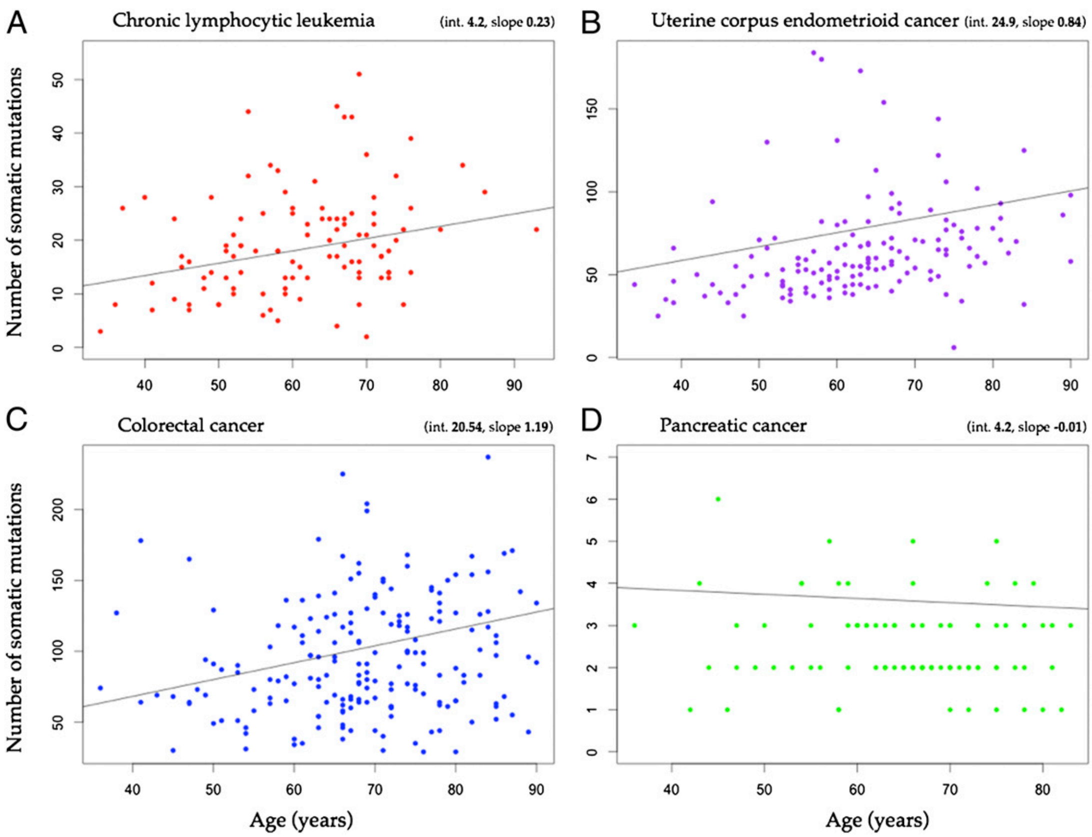

Fig. 2. Relationship between the number of somatic mutations and the age of diagnosis in four types of cancer: chronic lymphocitic leukemia (*A*), uterine cancer (*B*), colorectal cancer (*C*), and pancreatic cancer (*D*).

has self-renewed before tumor initiation. Using this formula, the slopes derived from the data regressions, and letting $S = 3 \times 10^7$ (whole-exome sequencing), the number of somatic mutations accumulated per base per year are estimated to be $7.67 \times 10^{-9} \pm 1.3 \times 10^{-9}$ (SE) in the normal lymphocytes that were precursors to CLL, and $3.97 \times 10^{-8} \pm 2 \times 10^{-9}$ (SE) in colorectal epithelial cells. Thus, by letting $t = 12$ per y in the normal lymphocytes that were precursors to CLL (estimated to divide approximately once a month; refs. 4, 16), and $t = 52$ per y in colorectal epithelial cells (about one renewal per week, ref. 3), we can estimate that the in vivo tissue-specific somatic mutation probability per base per cell division is $u = 6.4 \times 10^{-10} \pm 1.1 \times 10^{-10}$ (SE) in normal lymphocytes that were precursors to CLL, and $u = 7.6 \times 10^{-10} \pm 3.8 \times 10^{-11}$ (SE) in colorectal epithelial cells. These results are remarkably similar to the estimated mutation rates of normal cells and bacteria, obtained using a variety of other experimental techniques (2, 17–19) (*Supplementary Information*). Interestingly, our estimates of the somatic mutation rates in normal tissues are derived through a completely different approach—using somatic mutations in tumors rather than mutational data derived from the study of normal cells.

Discussion

In contrast to previous models, our mathematical model includes all relevant phases in which somatic mutations may accumulate in a tissue and by providing a way to estimate the background somatic mutation rate directly from sequencing data. Its predictions are validated by correlations between age and mutation number among patients with the same tumor type. In addition to the correlations described above, we found correlations between age and mutation number also in smaller datasets: glioblastoma (ref. 11, $P = 0.035$) and medulloblastoma (ref. 12, $P = 0.00027$). Similarly, a significant correlation was reported in neuroblastoma (10). In breast cancers, however, there was no correlation between number of mutations and age (20), $P = 0.33$ (estrogen receptor positive) and $P = 0.14$ (estrogen receptor negative), despite the fact that breast epithelial cells self-renew. It is possible that breast epithelial cell renewal is highly variable among individuals, given that it is dependent on hormonal status, number of pregnancies, breastfeeding history, etc. This would obscure any correlation between age of diagnosis and mutation number. Similarly, in ovarian high-grade serous adenocarcinoma (TCGA, 317 patients), we did not find a significant correlation ($P = 0.21$).

Strictly speaking, our model predicts a correlation with the number of tissue renewals rather than age per se. It is only when tissue renewal rates are relatively consistent among individuals that significant age vs. mutation correlations would be expected to exist.

In conclusion, our results suggest that in typical patients with cancers of self-renewing tissues, a large part of the somatic mutations occurred before tumor initiation. In CLL, colorectal, and

Table 1. Average number of somatic mutations in cancers of patients of various age

Tumor type	25-y-old	Median age at diagnosis	85-y-old
CLL	10.09 ± 1.29(SE)	18.53 ± 0.42(SE)	24.16 ± 1.03(SE)
Uterine cancer	45.96 ± 1.99(SE)	77.95 ± 0.60(SE)	96.48 ± 1.38(SE)
Colorectal cancer	50.2 ± 2.47(SE)	102.40 ± 0.69(SE)	121.38 ± 1.20(SE)

Table 2. Fraction of somatic mutations that originated before tumor initiation in patients of various age

Tumor type	25-y-old	Median age at diagnosis	85-y-old
CLL	(10.09 − 5.86)/10.09 = 0.42	(18.53 − 5.86)/18.53 = 0.68	(24.16 − 5.86)/24.16 = 0.76
Uterine cancer	(45.96 − 33.33)/45.96 = 0.27	(77.95 − 33.33)/77.95 = 0.57	(96.48 − 33.33)/96.48 = 0.65
Colorectal cancer	(50.2 − 50.2)/50.2 = 0	(102.4 − 50.2)/102.4 = 0.51	(121.38 − 50.2)/121.38 = 0.59

ovarian cancer patients of median age, half or more (68%, 57%, and 51%, respectively) of the passenger somatic mutations appear to have occurred before the tumor-initiating event.

These results have substantial implications for the interpretation of the large number of genome-wide cancer studies now being undertaken. They reinforce the idea that most somatic mutations observed in common adult tumors do not play any causal role in neoplasia; they in fact occurred in completely normal cells before initiation. They also indicate that patient age should be considered in statistical analyses of sequencing data. Sequencing data of younger patients' tumors may provide more reliable distinction of driver mutations by reducing the "noise" caused by the accumulation of passenger mutations occurring in normal tissues as individuals age.

Materials and Methods

In this section we provide a detailed description of our mathematical model as well as of the statistical analysis we performed. All relevant phases of a cancer tissue's history are included.

There are large differences among various types of tissues. In some tissues, there is a hierarchy among cells as well as a spatial organization. For example, the epithelial lining of the colon is divided into $\sim 10^8$ crypts, each maintained by stem cells that reside at the crypt base. In other tissues, there is no evidence of a hierarchical organization and the spatial structure may be quite fluid. We will derive formulas where cells with stem-like properties (asymmetric division, symmetric self-renewal, and differentiation) are considered (21). Wherever this assumption does not hold, the nonrelevant parameters should be set equal to zero. Note that in a tissue with a hierarchical structure, the focus of the analysis should be on the stem cells that maintain the tissue's homeostasis, because mutations in these cells will be transferred to all their progeny, whereas mutations occurring among the more differentiated cells will eventually be lost.

Development Phase. In this phase a precursor cell, derived from the zygote, undergoes a clonal expansion (typically in the fetus) from which the tissue under study is formed (the head of the fish). The mathematics for modeling this process has already been developed in Tomasetti et al. (22), where it is shown that the expected value for the total number T_i of cells with a mutation in a given nucleotide base i, present by the time the tissue is fully developed, is

$$E[T_i] = N\,u\,\log\,N\left(\frac{1-a/2-b}{1-a-2b-d/l}\right), \qquad [2]$$

where N is the total number of cells in the population (that is, in the fully developed tissue), u is the probability of a point mutation per base per cell division, a and b are the probabilities of asymmetric division and symmetric differentiation (possibly equal to 0), respectively, and d and l are the average cell death and division rates, respectively (ref. 22; Eq. **6**).

From Eq. **2**, it follows that the expected value for the total number X_D of point mutations found in a cell at the end of the development process is

$$E[T_i] = S\,u\,\log\,N\left(\frac{1-a/2-b}{1-a-2b-d/l}\right), \qquad [3]$$

where S is the total number of nucleotide bases sequenced.

Tissue Renewal Phase. Given the previously mentioned differences among tissues in their hierarchical and spatial organization, we will model two opposite scenarios and show that the resulting formulas are effectively equivalent.

Consider a tissue such as the colon where, say, there are a total of C crypts, with M stem cells per crypt (estimates found in the literature for the number of stem cells present in each colonic crypt vary from 5 to 60). Stem cells reside at the base of each colonic crypt. Given this spatial constraint, we can treat the evolutionary process occurring in each crypt independently. Assume that stem cells usually divide asymmetrically to maintain the crypt in equilibrium, except that when a stem cell dies it is replaced via symmetric self-renewal by another stem cell. It follows that the probability for a new point mutation to reach fixation within a crypt is given by $P^{FIX} = 1/M$. Because stem cells are long-lived, we approximate the process by disregarding the effect of deaths and consequent self-renewals on the number of mutational hits occurring in the crypt. Thus, strictly speaking, this process is not a Moran model because mutations do not occur only at self-renewal: here, the main (by approximation, the only) source of somatic mutations is given by asymmetric divisions. Take the average time between a stem cell's asymmetric divisions as the time unit. Then, the rate at which a given point mutation occurs in a crypt and reaches fixation within the crypt is

$$uP^{FIX}M = u. \qquad [4]$$

Thus, the timescale for a successful, fixated mutation is given by $1/u$. Also, if the average lifespan of a stem cell was the time unit, the expected amount of time it would take for this successful mutation to reach fixation in the crypt, T^{FIX}, i.e., conditional upon the event of fixation, can be calculated to be

$$E\left[T^{FIX}\middle|FIX\right] \approx -M\frac{(1-P^{FIX})}{P^{FIX}}\log(1-P^{FIX}), \qquad [5]$$

where FIX represent the event of fixation (A detailed proof can be found in Durrett, ref. 23, pp 48–50). Because $P^{FIX} \ll 1$, we can approximate the expression in Eq. **5** by M. Thus, letting c be the average number of asymmetric divisions occurring in the lifespan of a stem cell, the rate for the fixation process is approximately given by $1/(cM)$ and the timescale is cM. Therefore, given that the timescale for the fixation process is much smaller (i.e., faster) than the timescale for the occurrence of a successful mutation, i.e., $cM \ll 1/u$, it follows that we can treat each successful mutation hit independently from all others.

Take the average time between a stem cell's asymmetric divisions as the time unit. We then model the total number of asymmetric divisions occurring among stem cells in one crypt by a Poisson process $D(t)$ with rate M, and the total number of successful mutations in a crypt at a given nucleotide i up to time t, by the compound Poisson process $T_i(t)$,

$$T_i(t) = \sum_{j=1}^{D(t)} Y_j, \qquad [6]$$

where Y_j are independent Bernoulli random variables with mean equal to $uP^{FIX} = u/M$. Given that u is very small ($\sim 10^{-10}$), we can also regard $T_i(t)$ as the probability that the crypt has been hit by a successful mutation on base i by time t. Note that we could use different Y's for different nucleotide bases to allow for different mutation rates in different regions of the genome. From Eq. **6** it follows that $E[T_i(t)] = ut$. Let $X_S(t)$ be the total number of point mutations found in a randomly picked stem cell at time t (where time is measured from the start of the self-renewal phase), and let S be the total number of nucleotide bases sequenced, as before. By disregarding the possible mutations inherited from the development phase, and by noting that $X_S(t)$ is a sum of compound Poisson processes, we obtain

$$E[X_S(t)] = S\,u\,t. \qquad [7]$$

Note that this equation is similar to the one provided in ref. 2 but the model upon which it is predicated is stochastic rather than deterministic, and accounts for fixation and extinction of somatic mutations within the colonic crypt.

Consider now a tissue that, unlike the colon, has no hierarchy among the cell population and no spatial constrains. Let N be the total number of cells in the tissue. It can be shown that the fixation of a neutral point mutation never occurs, as here N is very large (precisely if $Nu > 1/2$). We can then consider the intermediate states, where each clone created by a neutral point mutation is independent of the other possible clones containing the same mutation. Let i be the total number of cells with a given base mutated. Then, we can write the following Kolmogorov forward equation for the Moran process (24):

$$\begin{aligned}\frac{\partial P(i,t|i_0,t_0)}{\partial t} &= P(i-1,t|i_0,t_0)\cdot\left(\frac{N-(i-1)}{N}(i-1)(1-u)+\frac{(N-(i-1))^2}{N}u\right)\\ &+P(i+1,t|i_0,t_0)\cdot\left(\frac{N-(i+1)}{N}(i+1)(1-u)+\frac{(i+1)^2}{N}u\right)\\ &-P(i,t|i_0,t_0)\cdot\left(2\frac{(N-i)}{N}i(1-u)+\frac{(N-i)^2}{N}u+\frac{i^2}{N}u\right),\end{aligned} \qquad [8]$$

where $P(i,t|i_0,t_0)$ is the probability that at time t there are i cells with a specific nucleotide base mutated, conditioned on having i_0 cells with that mutation at time t_0. The terms in the first row of the right-hand side represent the probability that from i – 1 mutated cells we get one more mutated cell, in one of following two ways: either due to the death of one of the N – (i – 1) wild-type cells and the division of a mutated cell, or due to the death of a wild-type cell followed by the division of a wild-type cell in which one of the two daughter cells gets hit by a mutation. Similarly, the second and third rows include the cases of going from i + 1 to i mutated cells in one step or staying in state i. We solve Eq. **7** by using either of the following diffusion approximations (23–25):

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[(1-2x)uP(x,t)]+\frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\frac{2(1-x)x}{N}P(x,t)\right], \qquad [9]$$

a Fokker–Planck parabolic partial differential equation, or

$$dX=(1-2X)\,u\,dt+\sqrt{\frac{2(1-X)X}{N}}dB, \qquad [10]$$

a stochastic differential equation. In both equations x represents the proportion of cells in the total population with the given mutation. Solving Eq. **9**, or Eq. **10**, with initial condition $x(0) = 0$ (i.e., no mutants at time 0), we obtain

$$E[X(t)]=\frac{1}{2}-\frac{e^{-2ut}}{2}, \qquad [11]$$

and because $ut \ll 1$ implies

$$e^{-2ut}\approx 1-2ut, \qquad [12]$$

then

$$E[X(t)]=ut. \qquad [13]$$

From Eq. **13** it follows that the expected value for the total number of point mutations found in a randomly picked cell at time t is

$$E[X_S(t)]=S\,u\,t, \qquad [14]$$

the same expression as in Eq. **7**, irrespective of the tissue hierarchical and spatial organization.

Importantly, this is also the expected number of point mutations originating in the tissue self-renewal phase and present in each one of the cancer cells, because the first driver mutation occurs in a cell within the healthy tissue and clonally expands to a population of cancer cells all containing the mutations present in that cell.

The last expression allows us to predict that the number of point mutations found in cancer tissues should correlate with the age of the patient, under the assumption that the time from tumor initiation to tumor detection is consistent among patients with a given tumor type. However, one critical issue is whether this correlation will be detectable in the data, given that if the amount of somatic mutations accumulating during tumorigenesis is much larger, the "signal" may get lost due to the unavoidable noise of the data. As we will see in *Phase Comparison*, our mathematical analysis actually predicts that a rather large component of the mutations originates during tissue renewal.

Tumor Formation Phase. Consider a tumor cell population generated by k sequential clonal expansions due to k driver mutations. For simplicity, we will not consider here the case of different multiple waves expanding simultaneously. Let v_j be the probability of a driver j mutation per base pair per cell division; d_j/l_j, the turnover rate of wave j, with $j\in 1,\ldots,k$; $\lambda_j = l_j - d_j$ the growth rate of wave j; $\alpha_j=\lambda_j/\lambda_{j+1}$; a_j and b_j are the probabilities of asymmetric division and symmetric differentiation (possibly equal to 0); and $K_j=(1-a_j/2-b_j)/(1-a_j-2b_j-d_j/l_j)$, a decreasing function of j, because fitness increases with each wave. Let s^j be the median time it takes for the j driver hit to occur, with $s^1 = 0$, and $N_j(t)$ is the population of wave j at time t. Then, it can be shown that (26), conditioned upon nonextinction,

$$E\left[N_1\left(s^2\right)\right]\approx\frac{1}{v_2}\left(\frac{\lambda_1}{\lambda_2}\right), \qquad [15]$$

and

$$E\left[N_j\left(s^{j+1}-s^j\right)\right]\approx\frac{\frac{1}{v_{j+1}}\left(\frac{\lambda_i}{\lambda_{j+1}}\right)}{\left(\frac{\pi\lambda_{j-1}/\lambda_j}{\sin\left(\pi\lambda_{j-1}/\lambda_j\right)}\right)^{\frac{\lambda_j}{\lambda_{j-1}}}}. \qquad [16]$$

By Eq. **2**, we can estimate the probability that the first cell hit by the kth driver has a mutation at a given base as

$$\sum_{j=1}^{k-1}u\log E\left[N_j\left(s^{j+1}-s^j\right)\right]K_j. \qquad [17]$$

Thus, the expected number of passenger mutations that are found in the first cell hit by the kth driver, and therefore common to the kth wave, is

$$\begin{aligned}E[X_C(k)]=Su\log\left(\frac{\alpha_1}{v}\right)K_1+Su\left(\log\left(\frac{\alpha_2}{v}\right)-\frac{1}{\alpha_1}\log\left(\frac{\pi\alpha_1}{\sin\pi\alpha_1}\right)\right)K_2+\cdots\\ +Su\left(\log\left(\frac{\alpha_{k-1}}{v}\right)-\frac{1}{\alpha_{k-2}}\log\left(\frac{\pi\alpha_{k-2}}{\sin\pi\alpha_{k-2}}\right)\right)K_{k-1},\end{aligned} \qquad [18]$$

where we have set all $v_j = v$ for simplicity.

Phase Comparison. We can now compare Eq. **3**, Eq. **14** (same as Eq. **7**), and Eq. **18**. Because the term Su is found in all those equations, we can focus on the other terms: $\log N\left(\frac{1-a/2-b}{1-a-2b-d/l}\right)$ for the development phase, t for the tissue self-renewal phase, and $\log\left(\frac{\alpha_1}{v}\right)K_1+\left(\log\left(\frac{\alpha_2}{v}\right)-\frac{1}{\alpha_1}\log\left(\frac{\pi\alpha_1}{\sin\pi\alpha_1}\right)\right)K_2+\cdots+\left(\log\left(\frac{\alpha_{k-1}}{v}\right)-\frac{1}{\alpha_{k-2}}\log\left(\frac{\pi\alpha_{k-2}}{\sin\pi\alpha_{k-2}}\right)\right)K_{k-1}$ for the tumor formation phase, to determine their relative importance.

During the development phase cells must divide mainly symmetrically (say, $a < 0.25$, $b = 0$) and death should be minimal (say $d/l<0.25$); then

$$9\le\log N\left(\frac{1-a/2-b}{1-a-2b-d/l}\right)\le 52.5, \qquad [19]$$

because $9 < \log(N) < 30$, for $10^5 < n < 10^{13}$.

Table 3. Estimated intercepts and slopes obtained from the four datasets

Tumor type	Intercept	Slope
CLL	4.22 ± 2.16 (SE), $P = 0.05$	0.23 ± 0.04 (SE), $P = 7.4 \times 10^{-11}$
Uterine cancer	24.91 ± 3.27 (SE), $P = 2.55 \times 10^{-14}$	0.84 ± 0.05 (SE), $P < 2 \times 10^{-16}$
Colorectal cancer	1.30 ± 5.53 (SE), $P = 0.81$	0.74 ± 0.08 (SE), $P < 2 \times 10^{-16}$
Pancreatic cancer	4.24 ± 1.13 (SE), $P = 0.00017$	–0.01 ± 0.017 (SE), $P = 0.38$

The number of times a tissue self-renews by time t is t, because time is measured with the average time between a cell's asymmetric divisions as the time unit. Depending on the type of tissue and on the age of the patient then, t may be close to zero or quite large, e.g., $t \sim 4{,}160$ in a colon of an 80-y-old person, because the tissue renews, on average, once a week.

For the tumor formation phase, by using $k = 10$, $v = 3.4 \times 10^{-5}$, $l = 0.5$, $d = 0.5$, and a selective advantage of $s = 0.04$ (Bozic et al., ref. 5), we can calculate that the term in parentheses in Eq. **18** is equal to 351.9 (a, $b = 0$ for simplicity). The result does not change much if we increase the number of drivers (with $k = 20$ drivers it is still < 500) or if we vary the terms inside the logarithm. The only sensitive parameter is K_j via the fitness advantage given by each successive driver. Thus, if we consider the smaller value $s = 0.004$ (5), then the same term in Eq. **18** becomes 3,177.

Comparing the numerical values we obtained for the different phases, it appears that the tissue renewal phase plays a key role in the accumulation of passenger mutations found in cancers of self-renewing tissues. For example, consider a 61-y-old CLL patient (median age at diagnosis in CLL). If leukemias take an average of 7 y to reach detection size (14), we can assume that this patient was a 54-y-old when the first driver mutation hit. If we use 3×10^7 for the number of bases sequenced in a cell (whole-exome sequencing), 5×10^{-10} as an estimate for the passenger somatic mutation rate (2, 5), and letting $t = 12$ divisions per y among hematopoietic stem cells (4, 16), then our model predicts (Eq. **1**) that this patient will have ~10 point mutations on average per cell when hit by the first driver $E[X_S(t)] = S\,u\,t = (3 \times 10^7 \times 5 \times 10^{-10} \times 54 \times 12) \approx 10$. Because the median number of somatic mutations found in the CLL dataset is 18, we then predict that half or more of the passengers originated in the precancer phase.

Statistical Analysis. We analyzed four whole-exome sequencing datasets publicly available on TCGA and the ICGC websites: CLL (ICGC-ISC/MICINN), uterine corpus endometrioid carcinoma (TCGA-UCEC), colorectal cancer (TCGA-COAD/READ), and pancreatic cancer (ICGC-JHU).

Kendall's correlation test is used so as not to enforce a linear positive correlation between age and number of mutations (as instead with Pearson's). Spearman's correlation test yields equivalent results. In the CLL dataset we removed 4 patients with more than 1,000 somatic mutations, given that all other 105 patients had less than 45 (if not removed, $P = 0.04$). In the CLL dataset we removed 4 patients with more than 1,000 somatic mutations, given that all other 105 patients had less than 45 (if not removed, $P = 0.04$). In the uterine dataset we removed 13 patients with more than 5,000 somatic mutations (if not removed, $P = 0.05$). In the colorectal dataset (7) we removed 34 patients whose tumors had between 300 and 20,000 somatic mutations, given that all other 190 patients had less than 250 (if not removed, $P = 0.0016$). For pancreatic cancer, we removed 24 samples having more than 30 mutations (the majority from cell-line derived data, if not removed, $P = 0.25$), given that all other 90 patients had less than 7.

To estimate a regression line for the mutation counts as a function of age (as depicted in Fig. 2), we used the robust generalized linear model approach implemented in the **glmrob** function in the **R** package **robustbase**. The distribution of counts at a given age is assumed to be Poisson, consistently with the conclusions of our mathematical model (*Supplementary Information*). Thus, the link function used in the Poisson regression is the identity. No patient (outlier) was excluded from the analysis. The use of a robust method provides a principled way to down-weigh individuals who have aberrantly high mutation rates compared with the Poisson distribution.

The resulting estimates for intercepts and slopes, using the robust generalized linear regression, are shown in Table 3.

ACKNOWLEDGMENTS. C.T. was supported in part by the National Institutes of Health (NIH) under Grant T32 CA009337. G.P. was supported in part by the NIH/National Cancer Institute Grant 5P30 CA006516-46.

1. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458(7239): 719–724.
2. Jones S, et al. (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci USA* 105(11):4283–4288.
3. van der Flier LG, Clevers H (2009) Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu Rev Physiol* 71:241–260.
4. Kiel MJ, et al. (2007) Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU. *Nature* 449(7159):238–242.
5. Bozic I, et al. (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA* 107(43):18545–18550.
6. Beerenwinkel N, et al. (2007) Genetic progression and the waiting time to cancer. *PLOS Comput Biol* 3(11):e225.
7. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.
8. Yachida S, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467(7319):1114–1117.
9. Welch JS, et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150(2):264–278.
10. Molenaar JJ, et al. (2012) Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* 483(7391):589–593.
11. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321(5897):1807–1812.
12. Parsons DW, et al. (2011) The genetic landscape of the childhood cancer medulloblastoma. *Science* 331(6016):435–439.
13. Bignell GR, et al. (2010) Signatures of mutation and selection in the cancer genome. *Nature* 463(7283):893–898.
14. Bizzozero OJ, Jr., Johnson KG, Jr., Ciocco A (1966) Radiation-related leukemia in Hiroshima and Nagasaki, 1946-1964. I. Distribution, incidence and appearance time. *N Engl J Med* 274(20):1095–1101.
15. Little MP (2009) Cancer and non-cancer effects in Japanese atomic bomb survivors. *J Radiol Prot* 29(2A):A43–A59.
16. Bradford GB, Williams B, Rossi R, Bertoncello I (1997) Quiescence, cycling, and turnover in the primitive hematopoietic stem cell compartment. *Exp Hematol* 25(5): 445–453.
17. DeMars R, Held KR (1972) The spontaneous azaguanine-resistant mutants of diploid human fibroblasts. *Humangenetik* 16(1):87–110.
18. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65(18):8111–8117.
19. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88(16):7160–7164.
20. Stephens PJ, et al.; Oslo Breast Cancer Consortium (OSBREAC) (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486(7403):400–404.
21. Morrison SJ, Kimble J (2006) Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature* 441(7097):1068–1074.
22. Tomasetti C, Levy D (2010) Role of symmetric and asymmetric division of stem cells in developing drug resistance. *Proc Natl Acad Sci USA* 107(39):16766–16771.
23. Durrett R (2008) *Probability Models for DNA Sequence Evolution* (Springer, New York).
24. Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
25. Gardiner CW (2009) *Stochastic Methods: A Handbook for the Natural and Social Sciences* (Springer, Berlin).
26. Durrett R, Moseley S (2010) Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor Popul Biol* 77(1):42–48.