Supplementary Information

August 19, 2019

1 Overview of the model

We simulate the process of tumorigenesis in a tissue $\mathcal{T}$ by simulating the evolutionary dynamics of all cells composing that tissue. Since the evolution of a tissue is mainly driven by the division and death rates of each stem cell present in that tissue, we let $\mathcal{T} := \mathcal{T}_s \cup \mathcal{T}_d$, where $\mathcal{T}_s$ is the set of stem cells and $\mathcal{T}_d$ is the set of non-stem cells, i.e. progenitors and fully differentiated cells. The key defining characteristic of what we define as stem cells is their ability to produce daughter cells that are as undifferentiated as their mother cell. We further partition the tissue, or equivalently the set of all its cells, into spatially separated units that we term "crypts". The evolution of the tissue has two phases: A development or growth phase, and a post-development or adulthood phase. The two phases transition at a certain time that we denote by T_G. During the development phase, the number of crypts forming the tissue will grow deterministically to then stay constant after T_G. Thus, when approaching T_G, the stem cells' high self-renewing division rate required to form that tissue will decrease as a function of time to a fixed value. This fixed value is the rate enabling the adult tissue to stay in homeostasis after time T_G.

The division cycle of each stem cell in our model can be described in the following way:

- Between any two events (division/death), each cell waits a random time according to an exponential distribution. The intensity parameter of that exponential distribution depends on the fitness of the stem cell.
- When an event occurs, the cell can either die or divide. The specific division and death rates of a stem cell will depend on the number of stem cells sharing the same crypt, the age of the individual (before T_G) (Note: how about afterwards, the aging effect), and the genotype of the cell, i.e, whether or not the cell is normal or mutated and, if mutated, depending on the type/number of driver mutations it has acquired.
- When the cell divides, three possible events can occur: 1) a symmetric division that creates two stem cell daughters, 2) an asymmetric division that creates one daughter stem cell and one more differentiated non-stem

www.pnas.org/cgi/doi/10.1073/pnas.1914589117

daughter cell, or 3) a symmetric differentiation where both daughters are non-stem cells.

- Once a non-stem cell is created, it goes through a differentiation cycle, i.e. a fixed number of N_{diff} divisions happening at deterministic time steps denoted by Δt_{prog}. After N_{diff} divisions, all the $2^{N_{\text{diff}}}$ fully differentiated cells descending from that original non-stem cell wait some deterministic length of time T_l and then die.

- Each crypt has a carrying capacity: The division rate decreases to a limiting rate as the total number of cells in a crypt increases.

- When a stem cell undergoes a symmetric division, any one of the two daughter cells may acquire a mutation with a certain probability. A mutation may also occur in the stem cell daughter during asymmetric differentiation. Note that we disregard the possibility of a mutation enabling a differentiated cell to undifferentiate and regain stem cell status. Therefore mutations in non-stem cells can be neglected since their effect will be lost.

- A driver mutation, defined as a mutation conferring a fitness advantage to the stem cell, can have three possible types of effects corresponding to the three different categories of genes that may be hit (see Fig. 7 in Vogelstein et al. Science 2013): Cell Fate (CF), Cell Survival (CS) and Genome Maintenance (GM). Hitting the CF gene category will result in an increase of the probability that a division, when it occurs, is symmetric, therefore decreasing by the same amount the probability of a symmetric differentiation or death (note: I have included death which is equivalent to symmetric differentiation for a stem cell). The probability of asymmetric division remains unchanged. Hitting the CS category will result in an increase in the overall division rate. Finally, hitting the GM category will result in increasing the probability of a mutation at division.

- It is possible to acquire multiple mutations in the same driver gene category, but we stop increasing the fitness advantage after two such mutations. Moreover, each category of genes is itself partitioned into pathways, and multiple mutations happening in the same pathway do not increase fitness after the first event.

- We declare that a tissue has cancer once we observe a clone of cells with more than $N_{c,\text{stem}}$ stem cells in a crypt whose total size, stem plus non-stem cells, must be at least $N_{c,\text{total}}$, and with a total of $N_{c,\text{mut}}$ driver mutations over at least $N_{c,\text{cat}}$ different gene categories. Here the $N_c = (N_{c,\text{stem}}, N_{c,\text{total}}, N_{c,\text{mut}}, N_{c,\text{cat}})$ defining cancer is tissue dependent. For example, for colon cancer, it is required to have a clone of cells with more than 64 stem cells in a crypt whose total size must be at least 10^8, and with a total of 3 driver mutations over at least 2 different gene categories.

2 Mathematical setting

2.1 Evolution of the number of crypts

The tissue is partitioned into crypts C_i, $i \in \{1, 2, 3, ..., N(t)\}$. The number of crypts $N_c(t)$ is a piecewise linear deterministic function of time. We start time ($t = 0$) at conception, with a first growth period until birth, which occurs at $t = 36$ weeks and that we define as time T_B (B for birth). For simplicity we make the approximation that after birth all tissues grow by scaling with the height of the individual; thus, a tissue reaches its full size when the individual reaches her/his final height, which we set at 20 years of age. The end of the growth period occurs at time T_H (H for homeostasis), with $T_H - T_B = 20$ years. The growth rate is constant between these times, null after T_H and three times faster before T_B than after this time. This faster rate is to fit the larger growth rate of a tissue during the fetus development. Letting N_c^0 be the target number of crypts, this leads to the following expression for the number of crypts at time t.

$$N_c(t) = \begin{cases} \left\lfloor 1 + 3\dfrac{N_c^0 - 1}{2T_B + T_H} t \right\rfloor & \text{if } 0 \le t \le T_B \\ \left\lfloor \dfrac{2N_c^0 T_B + T_H}{2T_B + T_H} + \dfrac{N_c^0 - 1}{2T_B + T_H} t \right\rfloor & \text{if } T_B \le t \le T_H \\ N_c^0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\lfloor x \rfloor$ denotes the integer part of x. During this growth phase, a crypt is created by duplicating a randomly selected existing crypt, therefore with exactly the same number of normal stem cells (which is fixed and equal to 10) and the same number of mutant stem cells for each genotype (see below). The number of normal stem cells in any crypt remains constant over time, even though additional stem cells will develop when mutations occur.

2.2 Stem cell genotype

The "genotype" of a stem cell v is summarized by three integers that represent the number of pathways that have been hit by a mutation in each of the three groups, CF, CS and GM, respectively, and we will write $g(v) = (\xi_1(v), \xi_2(v), \xi_3(v))$. For example, for a cell v, if CF has been hit in two different pathways, CS in one pathway and GM in zero, then $g(v) = (2, 1, 0)$. Normal cells have genotype $(0, 0, 0)$. The genotype determines the cell's fitness advantage. We do not increase this advantage beyond two hits in a given group, so that, for example, $g(v) = (3, 1, 0)$ induces the same fitness as $g(v) = (2, 1, 0)$. As, a consequence, we stop counting the number of hits beyond two, so that $\xi_j(v) = 2$ should be interpreted as the occurrence of two or more hits in different pathways in group j. We denote by $\mathcal{G}$ the set of possible "reduced genotypes" defined in this way, which, therefore, has 27 elements. We emphasize the fact that our model implicitly allows for multiple hits in the same pathway, or hits in more than two pathways, but our computation only keeps track of the reduced

information as defined above (and our transition probabilities are adjusted accordingly as described in section 2.6).

2.3 Division rate

The division rate of a cell v depends on the time t, genotype and the size of the crypt $C(v)$ that contains it. It is therefore driven by a function $\tau_{\text{div}} : \mathcal{G} \times \mathbb{N} \to \mathbb{R}^+$, so that the division rate of v is $\tau_{\text{div}}\left(g(v), |C(v)|, t\right)$, $|C(v)|$ being the total number of cells in $C(v)$. We first define a division rate for normal cells. We introduce a constant N_{cap} and we introduce the function τ_{div}:

$$\tau_{\text{div}}((0,0,0), N, t) = \left(M \exp\left(-\frac{1}{\left(1 - \frac{N}{N_{\text{cap}}}\right)^{\alpha}} \right) \mathbf{1}_{N<N_{\text{cap}}} + \tau_{\text{cap}} \right) a(t) \tag{2}$$

where:

$$a(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq T_H \\ \exp\left(-\log(2) \dfrac{t - T_H}{T_F - T_H} \right) & \text{if } t > T_H \end{cases} \tag{3}$$

where T_F is the final time of the simulation (75 years). The idea is that due to aging, the division rate starts decreasing after T_H, to attain half of the division rate of age T_H at time T_F. The constants τ_{cap} and α will be later adjusted based on constraints imposed on the model. As for the constant N_{cap}, we take it to be 10^9. Finally, the constant M is computed once all the other constants are fixed by imposing that $\tau_{\text{div}}((0,0,0), N, 0)$ is equal to the known normal division rate of the tissue. The division rate of a cell with genotype $g(v) = (\xi_1(v), \xi_2(v), \xi_3(v))$ is then defined by

$$\tau_{\text{div}}(g(v), |C(v)|, t) = c_{\text{app}}^{\xi_2(v)} \tau_{\text{div}}((0,0,0), |C(v)|, t), \tag{4}$$

where c_{app} is a constant driving the fitness advantage provided by mutations in the CS group and will be adjusted to fit the constraints.

2.4 Probabilities of symmetric division, symmetric differentiation and asymmetric division

For a given tissue, the probability of symmetric division p_{sym}, depends only on the (reduced) genotype of v, $g(v)$. More precisely, we define, for $g = (\xi_1, \xi_2, \xi_3)$

$$p_{\text{sym}}(g) = p_{\text{sym}}^0 + \xi_1 \delta, \tag{5}$$

where δ is a constant driving the fitness advantage provided by mutations in the CF group, and will be adjusted to fit the constraints. Similarly, the probability of symmetric differentiation $p_{\text{symdiff}}(g)$ is defined as:

$$p_{\text{symdiff}}(g) = p_{\text{symdiff}}^0 - \xi_1 \delta. \tag{6}$$

Finally, the probability of asymmetric division does not depend on the genotype and is defined as:

$$p_{\text{asym}} = 1 - (p^0_{\text{sym}} + p^0_{\text{symdiff}}) \tag{7}$$

2.5 Death rate

The death rate of a cell v, τ_{death} is taken to be a constant independent from the genotype of the cell and the size of the crypt.

$$\tau_{\text{death}}(t) = \tau_{\text{div}}\left((0,0,0),10,t\right)(p^0_{\text{sym}} - p^0_{\text{symdiff}}). \tag{8}$$

We require the equilibrium equation to yield a constant death rate that corresponds to a death every 2 years. This requires $(p^0_{\text{sym}} - p^0_{\text{symdiff}})$ to be equal to that death rate divided by $\tau_{\text{div}}\left((0,0,0),10,t\right)$, the normal division rate of the tissue, which is known if we fix the tissue.

2.6 Mutation events

Recall that a mutation can occur when we have a cell division followed by a symmetric division or an asymmetric division. We denote the probability that one daughter cell from a cell v acquires a mutation after a symmetric division by $\mu(g(v))$, where $\mu(g)$ has the form:

$$\mu(g) = (1 + \xi_3 c_m)\mu, \tag{9}$$

where c_m is a fixed constant. Denote by π_1, π_2 and π_3 the fraction of driver mutation loci in CF, CS and GM relative to the total number of such loci in the three groups together.Then, the probability that one of the daughter cells has a mutation of the corresponding type is $\pi_i\mu(g(v))$.

Recall that we do not increase fitness for cells having three or more mutations in the same group, or two or more in the same pathway, and that we focus our computation on "effective" mutations (those that confer an advantage) and therefore do not track the others. We now describe how we compute transition probabilities on the reduced genotype discussed in section 2.2.

For $j = 1, 2, 3$, let K_j denote the number of pathways that belong to group j, and $\beta_{j1}, \ldots, \beta_{jK_j}$ the fraction of driver mutation loci in each pathway relative to the total number of loci in the group. The probability of an efficient mutation of type j is then, for $g = (\xi_1, \xi_2, \xi_3)$,

$$\tilde{\pi}_j(g) = \begin{cases} \pi_j \text{ if } \xi_j = 0 \\ \pi_j\Big(1 - \sum_{k=1}^{K_j} \beta^2_{jk}\Big) \text{ if } \xi_j = 1 \\ 0 \text{ if } \xi_j = 2 \end{cases}$$

To understand the second expression, note that the probability of an effective second mutation in type j given that the first mutation was in pathway k is

$\pi_j(1-\beta_{jk})$. The probability that the first mutation of type j is in pathway k given that there is one mutation of type j is β_{jk}. This implies that the probability of a second effective mutation in type j given the current genotype is given by

$$\pi_j \sum_{k=1}^{K_j} \beta_{jk}(1-\beta_{jk}) = \pi_j\Big(1-\sum_{k=1}^{K_j}\beta_{jk}^2\Big).$$

Finally, letting $\tilde{\pi}(g) = \tilde{\pi}_1(g) + \tilde{\pi}_2(g) + \tilde{\pi}_3(g)$, we can define the effective mutation rate associated to g by

$$\mu'(g) = \tilde{\pi}(g)\mu(g),$$

so that the probability for a mutation of type i given an effective mutation is equal to $\pi'_i(g) = \tilde{\pi}_i(g)/\tilde{\pi}(g)$.

When we have an asymmetric division, the probability that the daughter cell acquires a mutation is simply defined as $\mu(g)/2$, and the probability of an effective mutation is $\mu'(g)/2$. The definition of π' remains unchanged.

In our simulations, we made the simplifying assumption that each group has two pathways of equal sizes, resulting in the correction $1-\sum_{k=1}^{K_j}\beta_{jk}^2$ being equal to $1/2$.

3 Markov chain model for the evolution of the tissue

3.1 Definition of the chain

Label the crypts at a given time by $(C_i, i \in \{1,2,3,...,N(t)\})$. For each i and $g \in G$, let $N_g^i(t)$ denote the number of stem cells in the crypt C_i that have (reduced) genotype g. Recall that every progenitor cell divides in two identical differentiated cells after every time interval Δt_{prog}, provided that its ancestor stem cell did not differentiate more than N_{diff} times. To keep track of this process, we let $D_i(t)$ denote the age distribution of the differentiated cells in C_i, which can be stored as a finite family of pairs (K, A) expressing that there are K differentiated cells in C_i with age K. We will let $X(t)$ denote the complete configuration

$$X(t) = \big(N(t), \big((N_g^i(t), g \in G), D^i(t), i = 1, \ldots, N(t)\big)\big)$$

4 Exact evolution of the Markov chain

4.1 Generation of the jump times

The rules determining the Markov chain transitions for a single cell are visualized in Figure 1, where possible transition types are represented as terminal nodes, labeled 1 to 10. A rate $\tau_{g(v)}^{(j)}$ equal to the product of the weights of the edges in

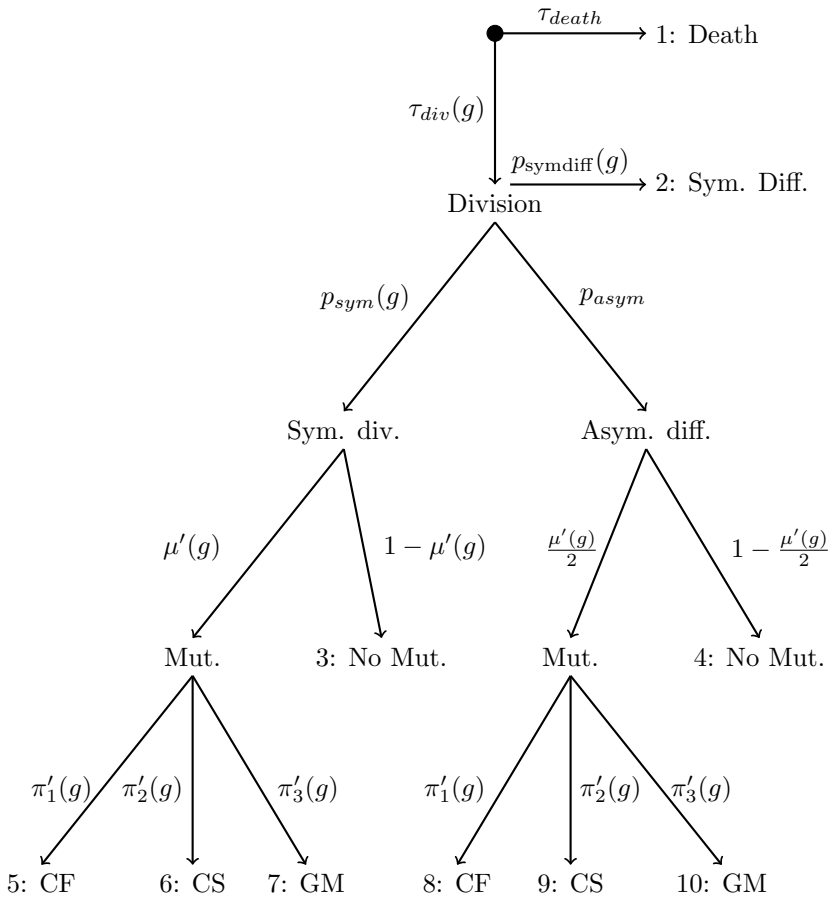

Figure 1: Visualization of the model's transition rules as a tree. Each terminal node of the graph represent a possible event, occurring with rate equal to the product of rates along the path linking the node to the root.

the path connecting a terminal node j with the tree root is attributed to each such node. Let us denote by $(T_k)_{k\geq 1}$ the stopping times when one of the stem cells of the tissue enters a division/death cycle. Then the times between events, $T_{k+1} - T_k$, are independent and have the same distribution as $\min_{j,v} U_{j,v}$ where the variables $(U_{j,v})$ are jointly independent and each $U_{j,v}$ follows an exponential distribution with rate $\tau^{(j)}_{g(v)}$. Furthermore, the transition occurring at that time is that corresponding to the cell v and index j achieving the minimum. The resulting change in the updated cell us described in Table 1.

Similarly, denote by $(T'_k)_{k\geq 1}$ the sequence of times where one or more differentiated cells divide or die, with $T_0 = T'_0 = 0$ and the chain starting from one crypt with 10 "normal" stem cells and 64 differentiated cells. This process is deterministic given the current state of the process at, say, time t, i.e., it is the first time after t for which one of the differentiated cells will reach an age equal to a multiple of Δt_{prog}. At such times, a cell whose age A is divisible by Δt_{prog} is duplicated if $A/T'_k < N_{\text{diff}}$ and removed if $A/T'_k = N_{\text{diff}}$, updating the distributions D_i accordingly.

5 Approximation and simulation

An exact simulation of the Markov chain as described above is, however, computationally intractable. Instead, we define a time step h and update the chain at times jh where j is an integer. Since the states of different crypts are independent, we focus on particular crypt C_l and describe how we update the chain from time jh to time $(j+1)h$.

The differentiated cells in C_l are first updated as follows. For an age a, let

$$m_a = \left\lfloor \frac{a+h}{\Delta t_{\text{prog}}} \right\rfloor - \left\lfloor \frac{a}{\Delta t_{\text{prog}}} \right\rfloor$$

denote the number of progenitor divisions occurring between a and $a+h$. Let $D^a_l(t)$ denote the number of progenitor cells in C_l with age a at time t. Because differentiated cells are created at times multiple of h and die after N_{diff} divisions, we need to keep track of the family

$$(D^{qh}_l(jh), q = 0, \ldots, k_h)$$

where $k_h = \lfloor N_{\text{diff}} \Delta t_{\text{prog}}/h \rfloor$. This family can be updated using

$$D^{(q+1)h}_l((j+1)h) = 2^{m_a} D^{qh}_l(jh) \text{ if } q+1 \leq k_h.$$

The number of new progenitor cells, $D^0_l((j+1)h)$ is determined from the stem cell evolution that we now describe.

Each event corresponding to terminal nodes $i = 1, \ldots, 10$ in Figure 1 is simulated in bulk. We approximate the number of cells with genotype g that experience this transition between times jh and $(j+1)h$ by the realization of a random variable $Y^{(i)}_g$ that follows a Poisson distribution with rate $hN^l_g(jh)\tau^{(i)}_g$

Node	Transition
1	$N^i_{g(v)} \to N^i_{g(v)} - 1$
2	$N^i_{g(v)} \to N^i_{g(v)} - 1$ and two differentiated cells are created in C_i with age 0
3	$N^i_{g(v)} \to N^i_{g(v)} + 1$
4	One differentiated cell is created in C_i with age 0
5	$N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v) + 1, g_2(v), g_3(v))$
6	$N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v), g_3(v) + 1, g_3(v))$
7	$N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v), g_2(v), g_3(v) + 1)$
8	One differentiated cell is created in C_i with age 0, $N^i_{g(v)} \to N^i_{g(v)} - 1$ and $N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v) + 1, g_2(v), g_3(v))$
9	One differentiated cell is created in C_i with age 0, $N^i_{g(v)} \to N^i_{g(v)} - 1$ and $N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v), g_3(v) + 1, g_3(v))$
10	One differentiated cell is created in C_i with age 0, $N^i_{g(v)} \to N^i_{g(v)} - 1$ and $N^i_{g'(v)} \to N^i_{g'(v)} + 1$ where $g'(v) = (g_1(v), g_2(v), g_3(v) + 1)$

Table 1: State transition in the Markov chain when cell v enters one of the 10 events associated with the terminal nodes in Figure 1.

and update the population numbers according to the transition type. More precisely, we let

$$\begin{aligned} N_g^l((j+1)h) =& N_g^l(jh) - Y_g^{(1)} - Y_g^{(2)} + Y_g^{(3)} + Y_{g_1'}^{(5)} + Y_{g_1'}^{(6)} + Y_{g_3'}^{(7)} \\ & - Y_g^{(8)} + Y_{g_1'}^{(8)} - Y_g^{(9)} + Y_{g_2'}^{(9)} - Y_g^{(10)} + Y_{g_3'}^{(10)} \end{aligned} \tag{10}$$

with the following notation: letting $g = (\xi_1, \xi_2, \xi_3)$, we set $g_1' = (\xi_1 - 1, \xi_2, \xi_3)$, $g_2' = (\xi_1, \xi_2 - 1, \xi_3)$ and $g_3' = (\xi_1, \xi_2, \xi_3 - 1)$ and $Y_{g'}^{(i)} = 0$ if one of the components of g' is negative. Moreover, if the right-hand side of (10) is negative, we set $N_g^l((j+1)h) = 0$

6 Free parameters: Tuning the parameters to fit constraints

6.1 Setting the goal

The parameters c_{app} in equation 4 (which drives the fitness advantage given by hitting the CS category), δ in equation 5 (which provides the fitness advantage given by hitting the CF category), α in equation 2 (which controls how fast the division rate decreases when the number of cells in a crypt augments) and τ_{cap} in equation 2 (which defines the division rate for a normal cell when the size of the containing crypt is larger than N_{cap}) are four free parameters in the model and are adjusted in order that the model fits an identical number of constraints.

Denote $\boldsymbol{\theta} = (c_{app}, \delta, \alpha, \tau_{cap})$, that we want to calibrate using macroscopic (population-level) data. Our constraints will take the form of known values of incidence of events associated with cancer or its progression, such as, for example, the risk of cancer occurring before a fixed age a in the population, or the probability of existence of a large crypt before age a.

In our model, these constraints are described by expectations $E(f)$ where f is a binary variable (taking values in $\{0, 1\}$) that depends on the whole trajectory of the Markov chain X between time 0 and some observation time, T. If we have c unknown parameters ($c = 4$ in our case), we need, in principle, c such functions, say $f_1, f_2, \ldots, f_c$, and we will tune $\boldsymbol{\theta}$ so that, for $i \in \{1, 2, ..., c\}$, $E(f_i)$ is equal to a known target, say γ_i.

We also enforce box constraints for $\boldsymbol{\theta}$ in the form $\boldsymbol{\theta} \in \mathcal{R} = \prod_{i=1}^{c} [a_i, b_i]$, where the a_i's and b_i's correspond to bounds for each of the parameters. So our calibration problem requires the solution of the system

$$\begin{cases} E\left(f_i|\boldsymbol{\theta}\right) = \gamma_i, \ \ i \in \{1, 2, ...c\} \\ \boldsymbol{\theta} \in \mathcal{R} \end{cases} \tag{11}$$

that we solve by minimizing

$$\sum_{i=1}^{c} \frac{\left(E\left(f_i|\boldsymbol{\theta}\right) - \boldsymbol{\gamma}_i\right)^2}{\boldsymbol{\gamma}_i} \tag{12}$$

over $\boldsymbol{\theta} \in \mathcal{R}$.

6.2 Gaussian process prior

Because an analytical computation of $E\left(f_i|\boldsymbol{\theta}\right) = \mathbb{P}\left(f_i = 1|\boldsymbol{\theta}\right)$ is intractable, we take a Bayesian optimization approach to address the problem [2]. For $i \in \{1, 2, ..., c\}$, we denote by $\phi_i(\boldsymbol{\theta})$ the function such that:

$$\mathbb{P}\left(f_i = 1|\boldsymbol{\theta}\right) = \frac{1}{1 + \exp\left(-\phi_i(\boldsymbol{\theta})\right)}. \tag{13}$$

For each i we assume a Gaussian Process (GP) prior distribution for $\phi_i(\boldsymbol{\theta})$ with mean 0 and a covariance kernel that we define below. More precisely:

$$\phi_i(\boldsymbol{\theta}) \sim GP(0, K_{\sigma,\boldsymbol{\rho}}(.,.)), \tag{14}$$

where:

$$K_{\sigma,\boldsymbol{\rho}}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = \sigma^2 \left(1 + \sqrt{3\sum_{j=1}^{4} \frac{\left(\boldsymbol{\theta}_{1j} - \boldsymbol{\theta}_{2j}\right)^2}{\boldsymbol{\rho}_j^2}}\right) \exp\left(\sqrt{3\sum_{j=1}^{4} \frac{\left(\boldsymbol{\theta}_{1j} - \boldsymbol{\theta}_{2j}\right)^2}{\boldsymbol{\rho}_j^2}}\right) \tag{15}$$

is a Matérn kernel with a fixed ν-parameter that is equal to $3/2$.

The sample paths of such a Gaussian process are almost surely one-time differentiable. We will solve problem (12) iteratively by sampling realization of the Markov chain with suitably chosen parameters. More precisely, assume that, after n steps, one has sampled n realizations of the Markov chain, with corresponding parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n$. Denote by f_i^l, $i \in \{1, 2, ..., c\}, l \in \{1, 2, ..., n\}$ the output (0 or 1) of the lth simulation (using the parameter $\boldsymbol{\theta}_l$) relative to the ith constraint. Fix a constraint (say i) and use for short the notation $\boldsymbol{\theta}^{(n)} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n)$, $\phi_i(\boldsymbol{\theta}^{(n)}) = (\phi_i(\boldsymbol{\theta}_1), \ldots, \phi_i(\boldsymbol{\theta}_n))$ and $f_i^{(n)} = (f_i^1, \ldots, f_i^l)$. The joint likelihood

$$\mathcal{L}(\phi_i(\boldsymbol{\theta}^{(n)}), f_i^{(n)}|\sigma, \boldsymbol{\rho})$$

is:

$$\left(\prod_{l=1}^{n} \left(\frac{\exp\left(-(1 - f_i^l)\phi_i(\boldsymbol{\theta}_l)\right)}{1 + \exp\left(-\phi_i(\boldsymbol{\theta}_l)\right)}\right)\right) \mathcal{N}(\phi_i(\boldsymbol{\theta}^{(n)})|0, \mathcal{K}_{\sigma,\boldsymbol{\rho}}(\boldsymbol{\theta}^{(n)})) \tag{16}$$

where $\mathcal{K}_{\sigma,\boldsymbol{\rho}}(\boldsymbol{\theta}^{(n)})$ is the matrix with coefficients $K_{\sigma,\boldsymbol{\rho}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, $i, j = 1, \ldots, n$. σ and $\boldsymbol{\rho}$ are estimated for each constraint i independently from the other constraints.

The covariance parameters σ and $\boldsymbol{\rho}$ are estimated at each step in order to maximize.

$$\mathcal{L}(f_i^{(n)}|\sigma, \boldsymbol{\rho}) = \int_{\mathbb{R}^n} \mathcal{L}(\psi_1, \ldots, \psi_n, f_i^{(n)} \,\big|\, \sigma, \boldsymbol{\rho})\, d\psi_1 \ldots d\psi_n. \tag{17}$$

We use Laplace's method to approximate this integral by

$$(2\pi)^{n/2} \frac{e^{m^*(\sigma,\boldsymbol{\rho})}}{\sqrt{\det(-H^*(\sigma, \boldsymbol{\rho}))}}$$

where $m^*(\sigma, \boldsymbol{\rho})$ is the maximum value (with respect to ψ) of $\log \mathcal{L}$ for fixed σ and ρ and $H^*(\sigma, \boldsymbol{\rho})$ is the Hessian of this function at the location $\boldsymbol{\psi}^{(n)}(\sigma, \rho) \in \mathbb{R}^n$ where this maximum is achieved. Similarly, the posterior likelihood:

$$\mathcal{L}_{\text{post}}(\phi_i(\boldsymbol{\theta}_1), \ldots, \phi_i(\boldsymbol{\theta}_n) \,\big|\, f_i^{(n)}, \sigma, \boldsymbol{\rho}) \tag{18}$$

is approximated as a Gaussian distribution $\mathcal{N}(\boldsymbol{\psi}^{(n)}(\sigma, \rho), -H(\sigma, \boldsymbol{\rho}))$ (using the just estimated σ and $\boldsymbol{\rho}$).

For an arbitrary parameter θ, we can compute the likelihood

$$\mathcal{L}(\phi_i(\theta)|f^{(n)}, \sigma, \rho) = \int_{\mathbb{R}^n} \mathcal{L}(\phi_i(\theta)|\phi_i(\theta^{(n)}) = (\psi_1, \ldots, \psi_n))\mathcal{L}_{\text{post}}(\psi_1, \ldots, \psi_n|f_i^{(n)})d\psi_1 \cdots d\psi_n \,,$$

which has closed form when $\mathcal{L}_{\text{post}}$ is approximated by a Gaussian. This computation allows us to determine the best next value of θ to perform a new simulation, by defining

$$\Psi_{n,i}(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{1}{1 + \exp\left(-\phi_i(\boldsymbol{\theta})\right)} \,\Big|\, f_i^{(n)}, \sigma, \boldsymbol{\rho}\right) \tag{19}$$

and minimizing

$$\min_{\boldsymbol{\theta} \in \mathcal{R}} \sum_{i=1}^{4} \frac{\left(\Psi_{n,i}(\boldsymbol{\theta}) - \boldsymbol{\gamma}_i\right)^2}{\boldsymbol{\gamma}_i}. \tag{20}$$

with respect to $\theta \in \mathcal{R}$.

In fact, we use a solution of (20) as the value of $\boldsymbol{\theta}^{(n+1)}$ only when n is even. When n is odd, we define $\boldsymbol{\theta}^{(n+1)}$ as a maximizer of

$$-\sum_{i=1}^{4} \left(\Psi_{n,i}(\boldsymbol{\theta}) \log \Psi_{n,i}(\boldsymbol{\theta}) + (1 - \Psi_{n,i}(\boldsymbol{\theta})) \log(1 - \Psi_{n,i}(\boldsymbol{\theta}))\right) \tag{21}$$

that measures the uncertainty of $\mathcal{L}(\phi_i(\boldsymbol{\theta})|f_i^{(n)}, \sigma, \boldsymbol{\rho})$. This allows us to explore new regions of the parameter space.

7 Description of parameters for each tissue

The complete set of parameters for colon is provided in table 2 (first column). The value of $p_{\text{asym}} = 0.9$ is justified in [5]. The values of p^0_{sym} and p^0_{symdiff} are deduced from the homeostasis equation

$$\tau_{\text{div}} p_{\text{sym}} = \tau_{\text{div}} p_{\text{symdiff}} + \tau_{\text{app}} \tag{22}$$

and the constraint $p_{\text{sym}} + p_{\text{symdiff}} + p_{\text{asym}} = 1$. All other parameters except c_{app}, δ, α and τ_{cap} (which are estimated using the approach described in section 6) are set based on biological information. In particular, the parameters π_1, π_2, π_3

describing the probabilities of mutation in groups CF, CS and GM are taken as $k\mu 10^{-3}$, where $\mu = 10^{-6}$ is the mutation rate over the whole genome and k is the number of genes in each group, as described in [3] for colon, [4] for blood and [1] for pancreas.

The parameters for FAP and Lynch syndrome are identical to those of the colon, with the following additional assumptions. For FAP, the probability of hitting CF for the first time is set to be 10^{-4} (instead of 10^{-6}), reflecting the loss of heterozygosity (rather than mutation) required for this event. For Lynch syndrome, we use a random mutation rate $\mu = X \times 10^{-6}$, where $X \sim \mathcal{N}(10, 5)$ is a Gaussian random variable with mean 10 and standard deviation 5. Also, we stop Lynch simulations at 50 years old.

Because FAP simulations create a very large number of clonal expansions that would be too computationally expensive to track over a large population, we have adopted the following approximation scheme, which uses the fact that, in our model, the 10^7 crypts of the tissue evolve independently. We simulate only one individual over a large period of time (150 years) and partition the tissue into 1,000 subsets of size 10^4 crypts. Denote by $T_1, T_2, ..., T_{1000}$ the times of cancer occurrence for each of the 1,000 subsets, so that $T = \min_i T_i$ provides the time of occurrence of cancer for the whole tissue. We can use our 1,000 simulations to estimate the distribution of T_i, which can be accurately modeled as a Gamma distribution, with known cumulative distribution function (c.d.f.) F_Γ. The c.d.f. for cancer occurrence, F_T is then given by

$$F_T(t) = 1 - (1 - F_\Gamma(t))^{1000}$$

and can be used to plot incidence curves.

For other tissues (blood and pancreas), for which p_{asym} is not known a priori, we determine p_{sym}, p_{symdiff} and p_{asym} based on (22) and the additional assumption that $p_{\text{sym}}/p_{\text{symdiff}}$ is tissue independent, so that this ratio can be deduced from that obtained in colon, which is $\sim$1.1 (and of course that these three probabilities sum to 1). In addition, the parameter c_{app} is estimated based on simulations, in order to match the lifetime incidence rate for each cancer. The corresponding parameters are listed in table 2.

8 Number of (passenger) mutations observed in a cell lineage

8.1 Setting

We here justify the formula used in the estimation of the fitness advantage arising from mutations observed in the CS group. This formula estimates the advantage through the ratio of the number of mutations per cell observed in cancel data to the number of mutations per cell that is expected in normal tissues. We first justify the expression we used for the latter quantity.

Parameters	Colon	Blood	Pancreas
Duration of simulation	75 years	75 years	75 years
Detection size	1.25×10^8 cells	1.25×10^8 cells	1.25×10^8 cells
N_c^0 (section 2.1)	10^7	10^7	2×10^8
T_B (section 2.1)	36 weeks	36 weeks	36 weeks
T_H (section 2.1)	20 years	20 years	20 years
τ_{div} (section 2.3)	1.75	0.219	0.029
N_{cap} (section 2.3)	10^9	10^9	10^9
α (section 2.3)	0.468 (*)	0.468	0.468
τ_{cap} (section 2.3)	0.413 (*)	0.413	0.413
c_{app} (section 2.3)	1.356 (*)	2 (*)	4.06 (*)
p_{sym}^0 (section 2.4)	0.0525	0.42	0.457
p_{symdiff}^0 (section 2.4)	0.0475	0.38	0.413
p_{asym} (section 2.4)	0.9	0.2	0.13
δ (section 2.4)	1.863×10^{-3} (*)	1.863×10^{-3}	1.863×10^{-3}
μ (section 2.6)	10^{-6}	10^{-6}	10^{-6}
c_m (section 2.6)	5	5	5
π_1 (section 2.6)	4×10^{-3}	7×10^{-3}	3×10^{-3}
π_2 (section 2.6)	7×10^{-3}	8×10^{-3}	11×10^{-3}
π_3 (section 2.6)	7×10^{-3}	10^{-3}	4×10^{-3}

Table 2: Parameters for colon, blood and pancreas. Values followed by (*) are adjusted numerically in order to fit population statistics relative to the relevant tissue. Others were either deduced from the literature or reproduced from values estimated for colon.

We focus on a single stem cell lineage and model the number of mutations observed over time. We assume that the cell divides with rate τ_{div}, and dies with rate τ_{death}. Cell division can be either symmetric (providing two stem cells) with probability p_{sym}, asymmetric (one stem cell and one progenitor) with probability p_{asym} or provide two progenitor cells with probability p_{symdiff}. It terms of lineage, the first two cases corresponds to survival, and the third one is equivalent to death. Finally, when a division leads to survival, it can lead to a mutation, which occurs with probability μ (otherwise the chain remains unchanged).

We assume that all acquired mutations are passengers in the sense that they will not affect the parameters of the model. We will be interested in the Markov Chain N_t providing the number of mutations at time t, with a special value denoted ω for the extinction of the lineage. We assume that $N_0 = 0$, and the possible changes of states for this are:

- $n \to \omega$ for any integer n with rate $\tau_- := \tau_{\text{death}} + \tau_{\text{div}} p_{\text{symdiff}}$
- $n \to n+1$ for any integer n with rate $\tau_+ := \tau_{\text{div}}(p_{\text{sym}} + p_{\text{asym}})\mu$
- ω is obviously a state from which the chain cannot escape.

We let T_e denote the extinction time, i.e., $T_e = \min\{t : N_t = \omega\}$. We also define $\tau := \tau_- + \tau_+$. With this notation, we now compute

$$\mathbb{E}(N_t | t < T_e).$$

8.2 Expected number mutations

First notice that $\mathbb{P}(t < T_e) = e^{-\tau_- t}$. Let $g(t) := \mathbb{E}(N_t 1_{t<T_e})$ and denote by T_0 the time of the first state change of chain. We have:

$$\begin{aligned} f(t) =& \mathbb{E}(N_t 1_{t<T_e}) \\ =& \mathbb{E}(N_t 1_{t<T_e} 1_{T_0>t}) + \mathbb{E}(N_t 1_{t<T_e} 1_{T_0 \le t}) \end{aligned}$$

Since $N_0 = 0$, the first term of the last equality vanishes. Using the strong Markov property and the fact that T_0 is exponential with rate τ, we have that:

$$\begin{aligned} \mathbb{E}(N_t 1_{T_0 \le t < T_e}) &= \int_0^t \tau e^{-\tau s} \left(\frac{\tau_+}{\tau}\right) \mathbb{E}((N_{t-s}+1) 1_{t-s<T_e}) ds \\ &= \int_0^t \tau e^{-\tau s} \left(\frac{\tau_+}{\tau}\right) \mathbb{E}(N_{t-s} 1_{t-s<T_e}) ds \\ &\quad + \int_0^t \tau e^{-\tau s} \left(\frac{\tau_+}{\tau}\right) \mathbb{P}(t-s < T_e) ds \\ &= \int_0^t \tau e^{-\tau s} \left(\frac{\tau_+}{\tau}\right) f(t-s) ds + \int_0^t \tau e^{-\tau s} \left(\frac{\tau_+}{\tau}\right) e^{-\tau_-(t-s)} ds \\ &= \tau_+ e^{-\tau t} \int_0^t e^{\tau u} f(u) ds + \tau_+ e^{-\tau t} \int_0^t e^{(\tau - \tau_-)u} ds \end{aligned}$$

In summary f is a solution of the ODE

$$f(t) = \tau_+ e^{-\tau t} \int_0^t e^{\tau u} f(u) ds + e^{-\tau t}(e^{\tau_+ t} - 1),$$

which can be solved explicitly yielding

$$f(t) = \tau_+ t e^{-\tau_- t}$$

and

$$\mathbb{E}(N_t | t < T_e) = \tau_+ t = t\tau_{\text{div}}(p_{\text{sym}} + p_{\text{asym}})\mu.$$

The formula used in the main paper is $\mathbb{E}(N_t|t < T_e) = t\tau_{\text{div}}\mu$ (or, $ad_T\mu$ when ages and rates are measured in years), therefore neglecting, for simplification, the factor $p_{\text{sym}} + p_{\text{asym}}$. Note that this factor is not important when computing the fitness advantage of mutations that only affect the division rate, since it cancels in the resulting ratio, as described below.

Returning to our model, the presence of a driver mutation in the CS group multiplies τ_{div} by c_{app} while leaving the other parameters unchanged. If such a mutation happens at $t = 0$, then the expected number of passenger mutations would be multiplied by c_{app}, so that this parameter represents, in our model, the fitness advantage $f = 1 + s$ provided by a mutation in CS. Still according to our model, a double mutation in CS gives $f = c_{\text{app}}^2$.

It is quite interesting that the values obtained for c_{app} in our model are qualitatively similar to the empirical ones listed in Fig. 2 of the main paper. Note, first, that the fitness advantages estimated from the observed number of passenger mutations at a given age is a lower bound of the fitness provided by the driver mutation, since the observed number of mutations includes a period of time during which the tissue is normal. Similarly, a tissue with two CS drivers mutations at a given age would have had a period of time with zero mutation, then one, then two.

Now, our simulations indicate that, according to our model, a very large majority of colon cancers involve a CS mutation occurring as a second driver, typically at age 35, with cancer being detectable about 35 years later. With $c_{\text{app}} - 1 \sim 0.3$, our model therefore predicts an empirical estimation of fitness of about 0.15, which is consistent with the near zero value reported in Fig. 2 of the main paper.

For blood, a first CS mutation is observed at roughly 15 years of age, with cancer declared about 40 years later. With $c_{\text{app}} = 2$, one would expect an estimated value of s equal to $40/55 \sim 0.7$, a value consistent with what is reported in the same Fig. 2.

Finally, for pancreas, the model requires two CS driver events, one roughly at 15, the second one 20 years later, with 30 more years before cancer is observed. Using our estimated value $c_{\text{app}} \sim 4$, our model would predict an empirical value of s equal to

$$s = \frac{15 + 4 \times 20 + 16 \times 30}{65} - 1 \sim 7.8\,.$$

This is smaller than the value reported in the main paper, which is 12, but definitely reflects a simlar trend compared to colon and blood cancers.

References

[1] Anirban Maitra and Ralph H Hruban. Pancreatic cancer. *Annu. Rev. pathmechdis. Mech. Dis.*, 3:157–188, 2008.

[2] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[3] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[4] The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[5] Cristian Tomasetti and Doron Levy. Role of symmetric and asymmetric division of stem cells in developing drug resistance. *Proceedings of the National Academy of Sciences*, 107(39):16766–16771, 2010.