Supporting Information

Ryser et al. 10.1073/pnas.1716552115

SI Text

Objectives and Assumptions

The process of colorectal tumorigenesis can be divided into three phases: progression from normal tissue to the first invasive tumor cell (transformation phase), expansion from the first tumor cell to the clinically detectable tumor mass at carrying capacity (growth phase), and tumor stasis characterized by an equilibrium between cell death and proliferation at carrying capacity (maintenance phase).

Our objective was to develop a spatial model of the growth phase that is consistent with knowledge about the biology of glandular homeostasis and carcinogenesis, that recapitulates our experimental findings, and that can be used to further investigate the role and importance of different biological aspects, such as the role of cellular mobility in shaping the observed spatial distributions of private mutations in the final tumor. The model development was based on the following biological assumptions:

i) In the absence of direct empirical evidence for cell death during the growth phase, we model the latter as an exponential clonal expansion without cell death. (Note that cell death is expected to be nonnegligible during the maintenance phase.)

ii) Each gland in the colon is populated by 8–32 stem cells at the gland bottom and up to 10,000 transit-amplifying and differentiated cells in the middle and upper sections of the gland. The stem cell population within each tumor gland is assumed to grow exponentially until the gland reaches a critical size, when fission (1) is initiated and the resident cell population is split between the two daughter glands.

iii) Pathology findings show that the organization of tumor cells within tumor glands (like in the normal colon) is conserved during growth of both adenomas and adenocarcinomas. The geometry of the tumor at carrying capacity appears to be complex, with substantial folding and intertwined glandular architecture. However, due to preservation of the glandular subunits during growth, and lateral fission of glands, the tumor is essentially 2D. The situation is similar to unfolding a piece of paper after crumpling it into a ball.

iv) Several studies have found neutral evolution to be the dominant regime for colorectal tumor growth (2, 3). Thus, we assume that the founding cancer cell has already acquired the phenotypic hallmarks of cancer (driver mutations). During the subsequent clonal expansion, only effectively neutral passenger mutations (not conferring any detectable selection) are acquired.

Two-Scale Model of the Expansion Phase

Based on these assumptions, we model the growth phase of colorectal carcinogenesis on two separate physical scales. On the larger scale we model tumor glands on a 2D lattice (Fig. S2*A*). At the finer scale of individual glands, we model the stem cell population as a cycle (Fig. S2*B*). The overall model structure is summarized by the following iterative growth dynamics:

- Step 0: Start of the founder gland.
- Step 1: Gland filling.
- Step 2: Cell mixing within the gland.
- Step 3: Gland fission.
- Go to Step 1, until final tumor size is reached.

Next, we describe the individual steps in more detail.

Step 0: Formation and Fission of the Founder Gland. The growth phase starts with a single cancer stem cell, which arises in the so-called founder gland (Fig. S2 *C*, *a*). In the framework of neutral evolution, the first cancer cell has acquired all necessary driver mutations, and hence it has a substantial growth advantage over the other stem cells in the founder gland, and it is assumed to replace them in a clonal expansion (Fig. S2 *C*, *b*).

Step 0-a: Gland filling. It is assumed that each cell has an exponential doubling time and hence that gland filling takes place in asynchronous fashion.

Step 0-b: Cell mixing within the gland. Once the founder gland is fully populated (Fig. S2 *C*, *c*), it undergoes mixing (Fig. S2 *C*, *d*) in the sense that cells with a mobile phenotype can exchange their position with a neighboring cell. In the process of mixing, cells can move to the opposite side of the gland. Cellular mobility is parameterized by the probability p of each stem cell to exchange its location with a neighboring stem cell ($p \in [0,1]$).

Step 0-c: Gland fission. Finally, having reached its carrying capacity, the founder gland undergoes fission and splits into two daughter cells that each carry half of the mother cell content (Fig. S2 *C*, *e*). Each of the two daughter cells and their respective progenies then repeat the following three steps.

Step 1: Gland Filling. New daughter glands are half-full after fission of the mother gland (Fig. S2 *C*, *e*). The resident stem cells now divide asynchronously and stochastically and push their neighboring cells along the growth direction to create space for their daughter cells, until the gland is filled to capacity (Fig. S2 *C*, *c*).

Step 2: Cell Mixing. Once the gland is full (Fig. S2 *C*, *c*), mobile cellular phenotypes can move around by exchanging their spot with a neighboring cell (Fig. S2 *C*, *d*). Each cell jumps with probability p. In the case of a jump, the cell exchanges its spot with one of the two nearest neighbors, chosen with equal probability 1/2. The order by which the stem cells jump is random. Importantly, through neighbor mixing, cells can end up on the "opposite side" of the gland (e.g., the bottom two cells in Fig. S2 *C*, *c* and *d*). By increasing the jump probability p from 0 (no jumps) to 1 (all cells jump), we model the phenotypic trait of mobility conferred by the driver mutations acquired before the exponential growth phase.

Step 3: Gland Fission. Once the gland has undergone mixing (Fig. S2 *C*, *d*), gland fission is initiated, and the gland splits according to the fission axis, which is aligned with the growth direction. One of the two daughter cells remains in the location of the mother gland, whereas the second daughter gland creates space through pushing other cells outward. We end up with two half-full daughter glands (Fig. S2 *C*, *e*), which themselves then undergo growth, mixing and fission, etc. Two important aspects about the fission process need to be discussed in detail: how space is created for the daughter cell that is placed on a neighboring spot (pushing) and synchronicity during early tumor growth.

Pushing. As a new gland is created during fission, space needs to be created in absence of cell death (as is the case during the expansion phase). We model the pushing process by choosing a pushing direction, generating a random path into that direction, and moving all cells along the path outward. See Fig. S3 for details about the pushing process.

Synchronicity. During the early stages of tumor growth, we assume that the time needed to regrow and split a gland is approximately deterministic and that gland fission events are largely synchronous.

For example, after fission of the founder gland, it is unlikely that daughter gland 1 and its own progeny all undergo fission before daughter gland 2 undergoes fission for the first time. After a number of fission events (set to 10th gland fission generation, or 1,024 glands), we relax this condition and assume asynchronous gland fission.

Mutation Accumulation

In the neutral evolution model, driver mutations have already been acquired once the exponential growth phase kicks in. Therefore, the tumor only accumulates neutral passenger mutations during its growth. Furthermore, because only early mutations are detectable in an exponentially growing population (Fig. 2*B*) and because there is a sizable number of private mutations found in excised tumor specimens (main text), there is a relatively high frequency of mutation accumulation during the first few cell divisions. For this reason, we assume that each new cell division is accompanied by a burst of mutations in each daughter cell, and the number of mutations acquired per burst is modeled as a Poisson random variable with mean λ. We keep track of the mutations acquired up to the fifth generation (32 cells), allowing for a total of 64 mutation bursts with an average of 64λ private mutations. The genotypes of cells and clonal glands are then determined based on the acquired private mutations.

Bulk and Gland Sampling in the Final Tumor

Tumor growth is simulated until the total size reaches 750,000 glands, or 6×10^6, 1.2×10^7, and 2.4×10^7 stem cells for scenarios with 8, 16, and 32 stem cells per gland, respectively. Assuming that there are 10,000 cells per tumor gland, the final tumors contain on the order of 10^{10} cells. To emulate the in vivo experimental design, we extracted from the final tumor two bulk samples in the form of two 25,000-gland patches [sampled from opposing regions of the tumor (Fig. 2*A*)]. Next, we determined the allele frequency of each private mutation in both samples and discarded mutations with allele frequencies below the threshold frequency of 10% (estimated sensitivity in next-generation sequencing experiments) in both bulk samples. For gland genotype analyses, we further sampled five glands from each bulk.

Model Constraints

In developing the final model as described above, we went through several iterations of informal model selection to ensure that the various model properties are consistent with experimental findings. Since these intermediate selection steps are no longer visible in the final model, and insights gained may be informative to the reader, we highlight here the key insights gained during the model development process.

First, the experimental observation that private mutations were side-specific in all adenomas (4/4) and more than half of carcinomas (9/15) imposed constraints on several modeling aspects. For instance, only models with little stochasticity and substantial geometric structure during early growth were able to recapitulate the observed private mutation patterns. For this reason, we introduced the notion of a growth direction during gland regrowth (Fig. S2), as well as the necessity for synchronous gland fission during the first 10 gland generations. Without these constraints, we found that side mixing of private mutations was too common even in tumors with no cellular motility ($p = 0$).

Second, the choice of a rectangular lattice to model the spatial arrangement of the glands was motivated by computational considerations and findings by ref. 4.

Third, the incorporation of mechanical aspects such as cell-cell forces and pressure gradients would have rendered the model more realistic but computationally intractable. Due to the rectangular lattice structure and the lack of pressure gradients, we introduced the random pushing path (Fig. S3) to obtain spherical tumors.

In summary, we developed a parsimonious neutral evolution model of colorectal carcinogenesis that is consistent with biologic knowledge and available data. A formal comparison of the early cell mobility model against plausible alternatives is found in *Bayesian Model Selection*.

ABC

To fit the model simulations to the targeted sequencing data from the human colorectal tumors, we used ABC, a framework that allows for approximate Bayesian inference in the absence of an analytic likelihood (5, 6). In essence, the model-based simulated data (D^*) and the experimental data (D) are compared based on a summary statistic (S) and a distance function $\rho(S(D^*),S(D))$. To this end, we first discretized the space of model parameters $\theta = (\lambda, n, p)$ as follows: the mutation burst rate λ was discretized into nine equal intervals (log-scale) over the range [0.05, 5.4]; the number of stem cells n was discretized (log-scale) into (8, 16, 32), and the cell mobility parameter p was discretized into four equal intervals over [0,1]. To capture salient features of the available data, we chose a multivariate summary statistic $S = (S_1, S_2,...S_k)$ where S_i were the rescaled (mean zero and SD 1) versions of the following:

- S_1: Fraction of unique gland genotypes in bulk A.
- S_2: Fraction of unique gland genotypes in bulk B.
- S_3: Number of private mutations in bulk A.
- S_4: Number of private mutations in bulk B.
- S_5: Side-mixing present (binary).
- S_6: Mean pairwise distance between glands in bulk A.
- S_7: Mean pairwise distance between glands in bulk B.

The distance function was defined as an L_2 distance

$$\rho(S(D^*), S(D)) = \left(\sum_{i=1}^{k} |S_i(D^*) - S_i(D)|^2 \right)^{1/2}.$$

The idea behind ABC is to introduce a threshold ε that discriminates between acceptable ($\rho \leq \varepsilon$) and unacceptable ($\rho > \varepsilon$) parameter values. We precomputed 100 simulations for each data point θ_i and used all simulations 15,000 simulated tumor samples to normalize the individual summary statistics and to compute ε (as the fifth percentile of ρ across all simulations). Then, assuming uniform prior distributions (on the log scales for λ and n), we applied the following rejection algorithm:

Step 1) Choose a parameter value on the grid θ_i, uniformly at random.

Step 2) Run the computational model to generate the model output M.

Step 3) Accept θ_j if $\rho(S(M), S(D)) < \varepsilon$, and reject otherwise. Return to step 1.

The marginal posterior distributions are shown in Fig. S4 for adenomas, nonmixing carcinomas, and mixing carcinomas. Posterior checks were performed for all tumor samples to ascertain satisfactory goodness of fit; see Fig. S5 for examples.

Bayesian Model Selection

We compared the performance of the model with early cell mobility (model M_1) against three a priori plausible alternatives, outlined below.

Model with Selection (M_2). Side mixing of private mutations could potentially be due to the presence of expansive clones that are selected for during tumor evolution. To model this scenario, we assumed that one of the early mutations (chosen uniformly at random) conferred a selective advantage s (i.e., the cell division rate of mutant cells is $1 + s$ times higher than the rate of cells without the driver mutation). Given the paucity of driver mutations found in each branch of the ancestral trees (Table 1), we limited the model to a single driver mutation in addition to the public drivers present in the first tumor cell. The selection strength of cancer driver mutations remains poorly characterized, but studies suggest that s is <10% (7, 8). We therefore explored a discretized fitness advantage of $s \in (0\%, 2.5\%, 5\%, 7.5\%, 10\%)$. The number of stem cells and mutation burst rate were modeled as in the early cell mobility model M_1. The posterior distributions for the model are shown in Fig. S7.

Delayed Cell Mobility Model (M_3). To account for the possibility that the cell mobility is not a trait present in the first tumor cell, but due to an additional mutation in the expanding clone, we allowed for a delay in cell mobility up to a tumor size of k glands with $k = 0,1,2,4,8$. The number of stem cells and mutation burst rate were modeled as in the early cell mobility model M_1. The posterior distributions for the model are shown in Fig. S8.

Self-Seeding of Tumor Cells (M_4). Previous work suggests that cancer cells may enter the circulation and reseed to other locations of the primary tumor (9). Because such a model of self-seeding tumor cells could potentially lead to a side-mixed mutation pattern, we explored an alternative model formulation with long-range seeding of tumor cells. To this end we introduced a self-seeding rate per gland fission r. We discretized r on a logarithmic scale as $(4 \times 10^{-6}, 1.4 \times 10^{-5}, 5.2 \times 10^{-5}, 1.8 \times 10^{-4}, 6.7 \times 10^{-4})$, which translates into an expected number of self-seeded tumor cells in the final tumor of (3.0, 11, 39, 139, 500). For each self-seeding event, a randomly selected tumor cell from the dividing mother gland was inserted in a tumor gland chosen uniformly at random from the entire tumor mass. The number of stem cells and mutation burst rate were modeled as in the early cell mobility model M_1. The posterior distributions for the model are shown in Fig. S9.

Model Selection. To estimate the posterior model probabilities $P(M_i|D)$ for each tumor, we then performed rejection sampling using a joint space-based approach (10). More precisely, starting from a noninformative prior $P(M_i) = 0.25$, $i = 1,2,3,4$, we implemented the following algorithm:

i) Draw a model M* according to the prior distribution;
ii) For model M*, draw a parameter set θ* based on the prior distributions $P(\theta|M^*)$;
iii) Simulate data M* with parameters θ*, yielding simulated tumor data D*;
iv) Compute the distance $\rho(S(D^*),S(D))$; accept if $\rho < \varepsilon$, and reject otherwise; and
v) Repeat.

Finally, the posterior distributions over the joint model and parameter space were marginalized to obtain the posterior (marginal) model probabilities (Fig. S6). As expected, for adenomas and nonmixing carcinomas, the data does not provide evidence for or against any particular model. For the mixing carcinomas however, the posterior distributions suggest that the early cell mixing model best recapitulates the data. The Bayes factors for the comparison of M_1 against M_2, M_3, and M_4, defined as $BF(M_1{:}M_k) = P(M_1|D)/P(M_k|D)$, ranged from 2 to 10.

1. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39:e118.
2. Sottoriva A, et al. (2015) A Big Bang model of human colorectal tumor growth. *Nat Genet* 47:209–216.
3. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A (2016) Identification of neutral tumor evolution across cancer types. *Nat Genet* 48:238–244.
4. Baker AM, et al. (2014) Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. *Cell Rep* 8:940–947.
5. Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 100:15324–15328.
6. Sottoriva A, Spiteri I, Shibata D, Curtis C, Tavaré S (2013) Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer Res* 73:41–49.
7. Waclaw B, et al. (2015) A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* 525:261–264.
8. Bozic I, et al. (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA* 107:18545–18550.
9. Kim MY, et al. (2009) Tumor self-seeding by circulating cancer cells. *Cell* 139:1315–1326.
10. Toni T, Stumpf MP (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26:104–110.

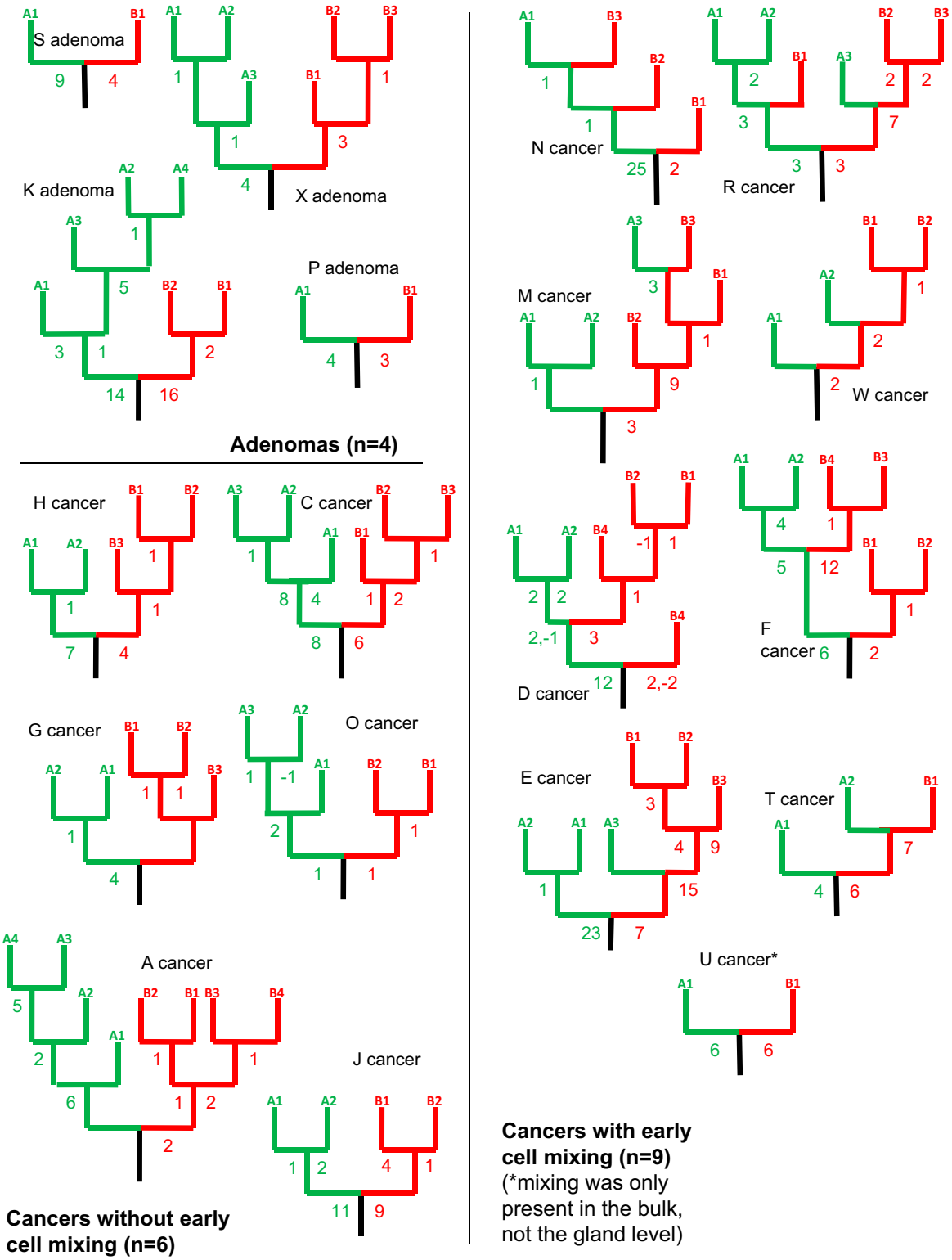


Fig. S1. Ancestral trees. For each of the 19 human colorectal tumors, the ancestral trees were reconstructed based on the mutations in sampled glands.

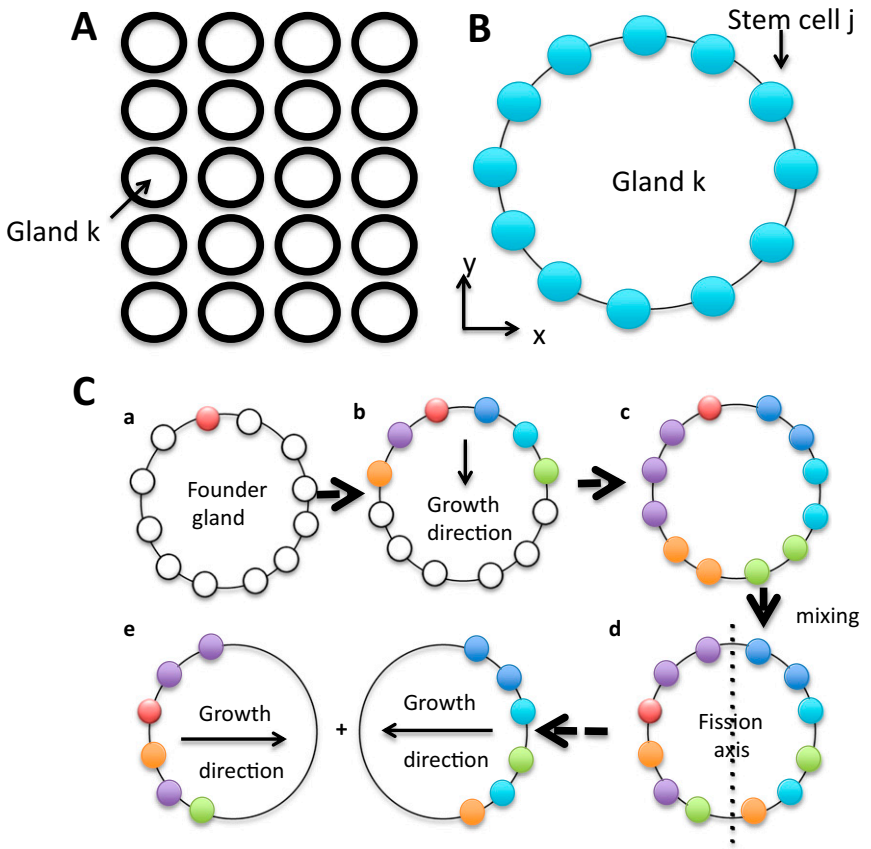


Fig. S2. Multiscale model of tumorigenesis. (*A*) Glands populate a 2D lattice. (*B*) Within each individual gland, the resident stem cells are positioned on a cycle. (*C*) The expansion phase starts with the first tumor stem cell (red) in the founder gland (*a*). Having acquired all driver mutations, this first cell outcompetes the precancerous cells (white) and fills the founder gland in the growth direction (*b*). Once the gland is full (*c*), the presence of cellular mobility enables mixing of the founder gland (*d*), which may cause cells to move to the opposite side of the fission axis (e.g., bottom two cells in *c* and *d*). Gland fission is then initiated (*d*) and leads to two daughter glands (*e*), whose growth directions are perpendicular to the fission axis of the mother gland. The two daughter cells now undergo the same cycle of growth, mixing, and fission.

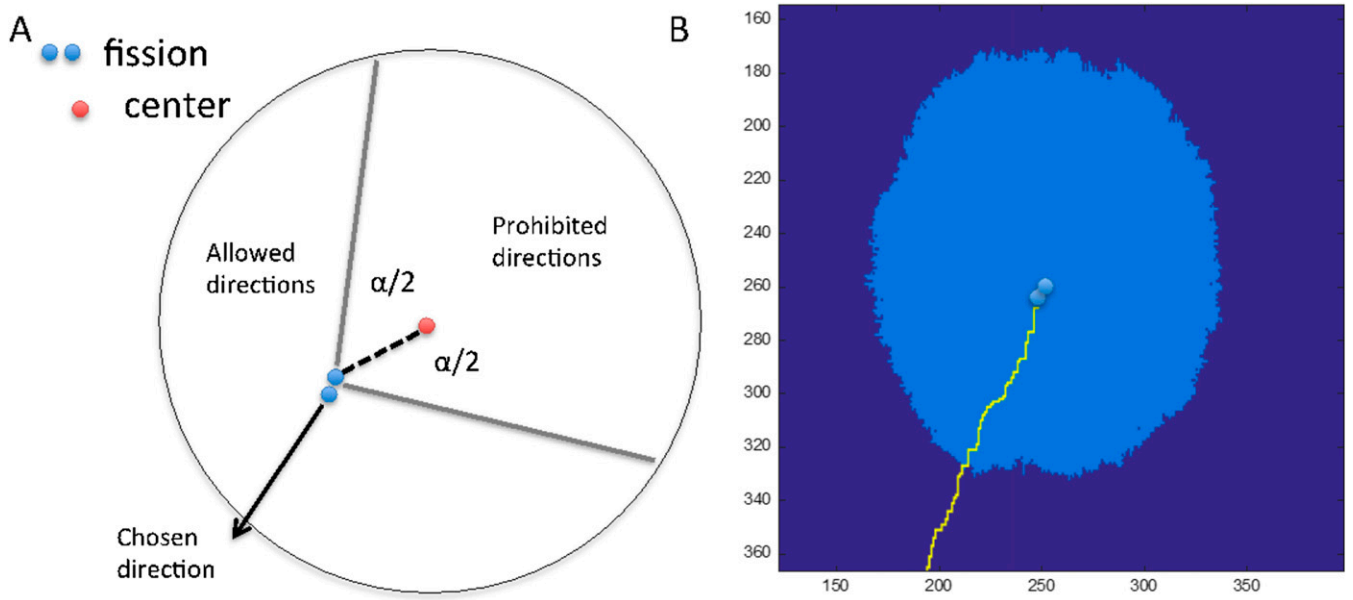
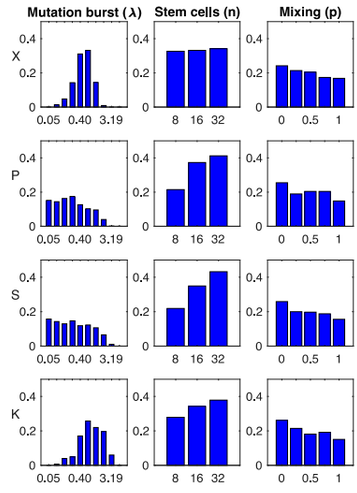


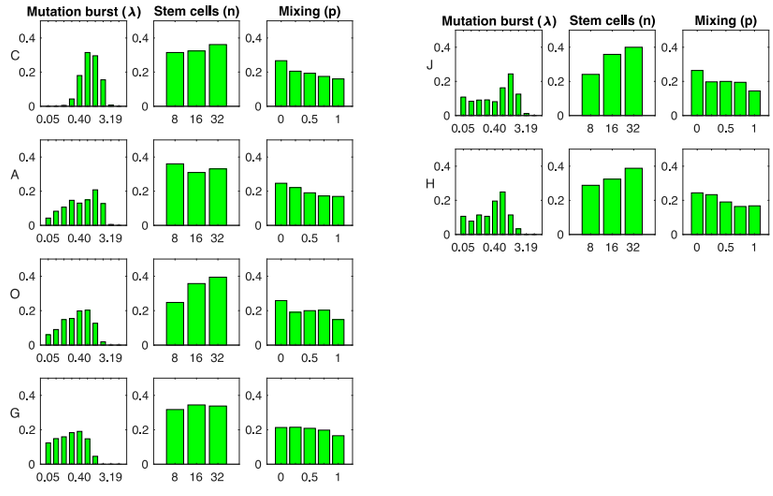
Fig. S3. Pushing dynamics after gland fission. (*A*) Upon fission of the mother gland, space needs to be created for the two daughter glands (blue dots). Connecting (dotted line) the location of fission with the tumor center (red dot), we mark a sector (angle α) as prohibited. A "pushing direction" is then selected among all allowed directions, uniformly at random. An angle of $\alpha = 90$ was sufficient as to ensure spherical tumors. (*B*) Along the pushing direction, we generate a stochastic path illustrated in yellow, and space for the additional cell is created by pushing resident cells outward along this path.

Model M_1
Early cell mobility

A - Adenomas



B - Non-mixing carcinomas



C - Mixing carcinomas

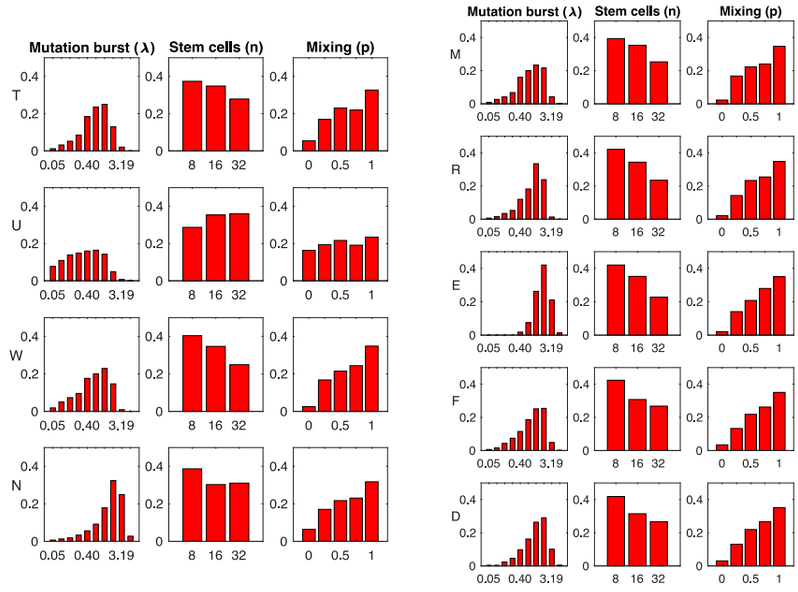


Fig. S4. Posterior parameter distributions for early cell mobility model. The posterior distributions for the mutation burst rate λ (log-scale), the number of stem cells *n* (log-scale), and the cellular mobility *p* are shown for four adenomas (*A*), six nonmixing carcinomas (*B*), and nine mixing carcinomas (*C*). The letter to the left indicates the tumor name (see also Table 1), and the *y* axis shows the posterior distribution of parameters.

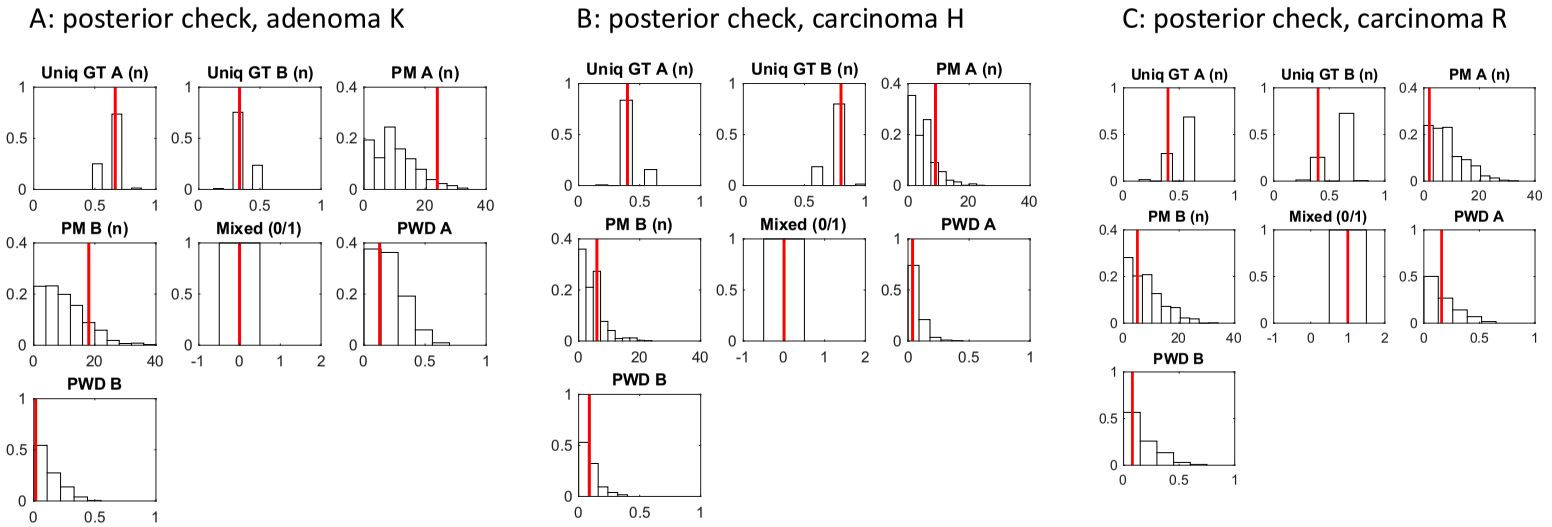


Fig. S5. Posterior checks. Posterior checks for adenoma K (*A*), nonmixing carcinoma H (*B*), and mixing carcinoma R (*C*) are shown for the seven components of the multivariate summary statistics. For each summary statistic, the posterior distribution (blue, *y* axis) is shown over the range of the standardized summary statistic (*x* axis). The value of the corresponding summary statistic for the experimental data are represented as a red vertical bar. Mixed, private mutations present on both tumor sides, yes/no; PM A/B, number of private mutations in bulk A/B; PWD A/B, mean pairwise distance between glands in bulk A/B; Uniq GT A/B, unique fraction of gland genotypes in bulk A/B.

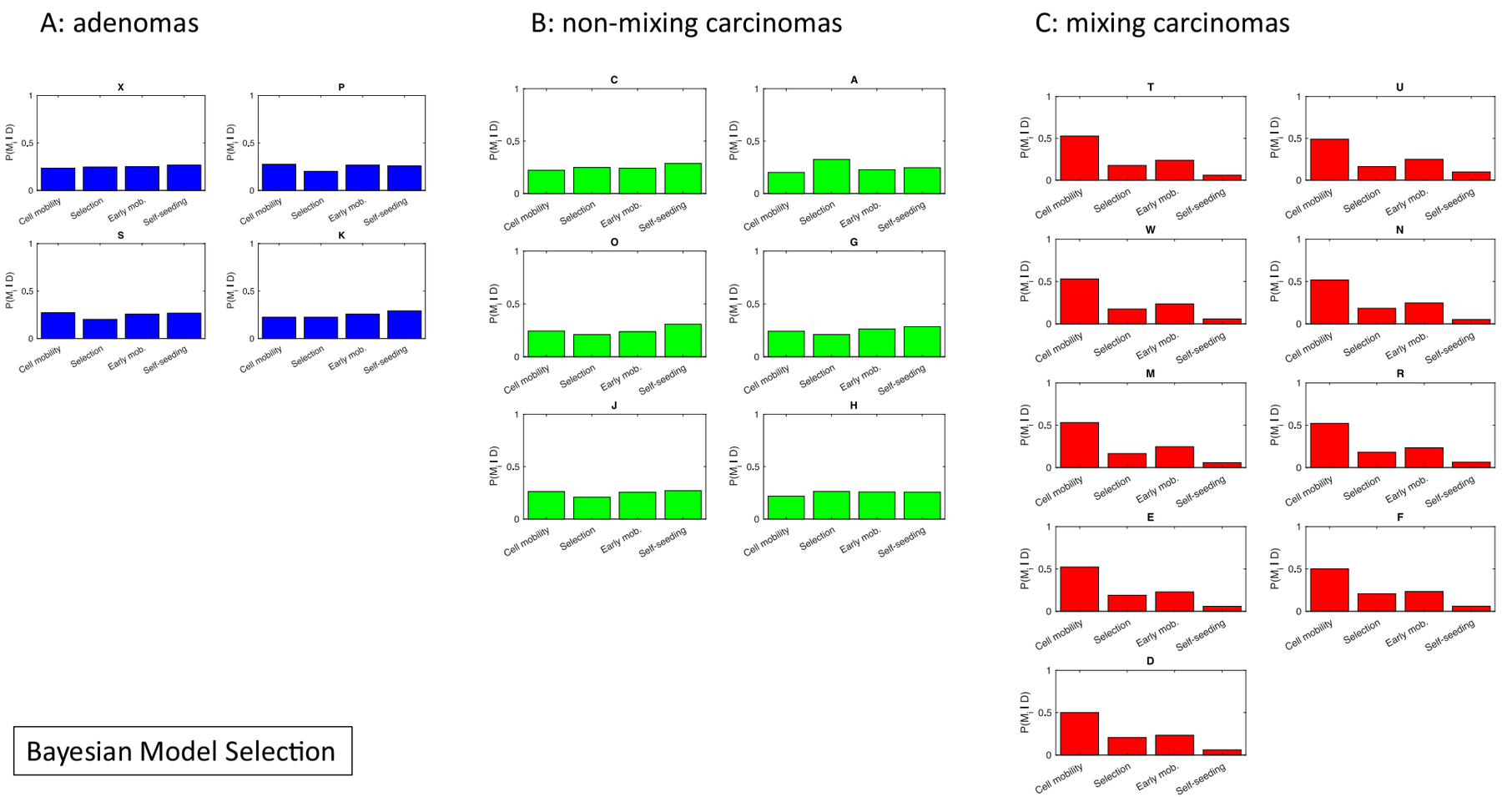
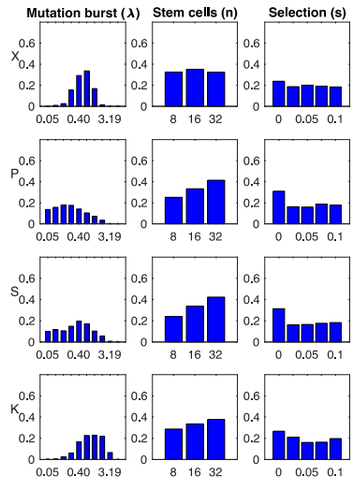
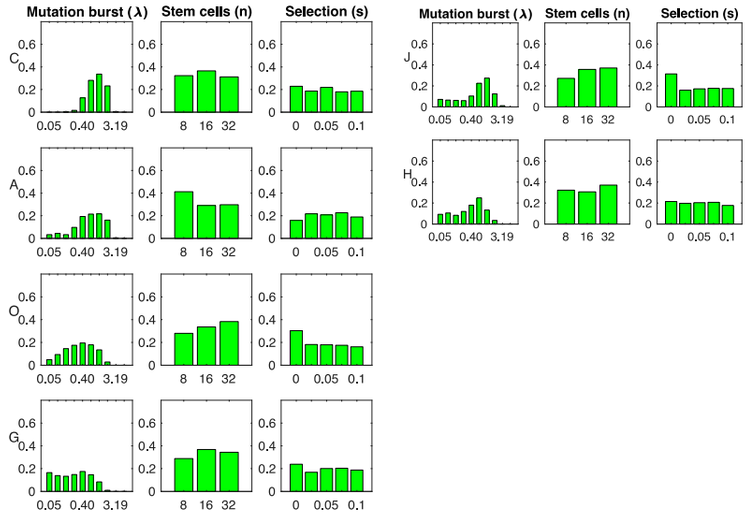


Fig. S6. Bayesian model selection. For each tumor, the marginal posterior probabilities for four models are shown. Model M_1, early cell mixing; model M_2, selection; model M_3, delayed cell mixing; model M_4, tumor cell self-seeding. The tumors are grouped into adenomas (*A*), nonmixing carcinomas (*B*), and mixing carcinomas (*C*). The marginal posteriors were calculated based on approximate Bayesian rejection sampling on the joint model-parameter space; see *SI Text* for details.

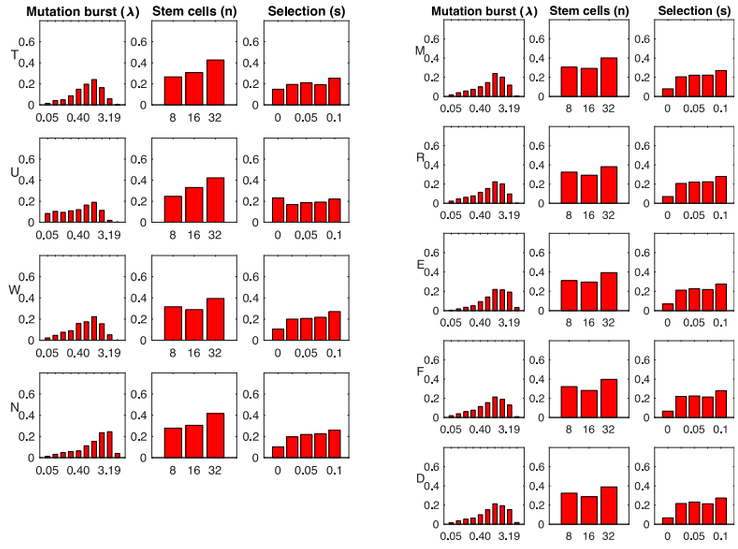
Model M_2
Selection



A - Adenomas



B - Non-mixing carcinomas



C - Mixing carcinomas

Fig. S7. Posterior parameter distributions for selection model. The posterior distributions for the mutation burst rate λ (log-scale), the number of stem cells *n* (log-scale), and the selection strength *s* are shown for four adenomas (*A*), six nonmixing carcinomas (*B*), and nine mixing carcinomas (*C*). The letter to the left indicate the tumor name (see also Table 1), and the *y* axis shows the posterior distribution of parameters.

Model M_3
Delayed mixing

A - Adenomas

B - Non-mixing carcinomas

C - Mixing carcinomas

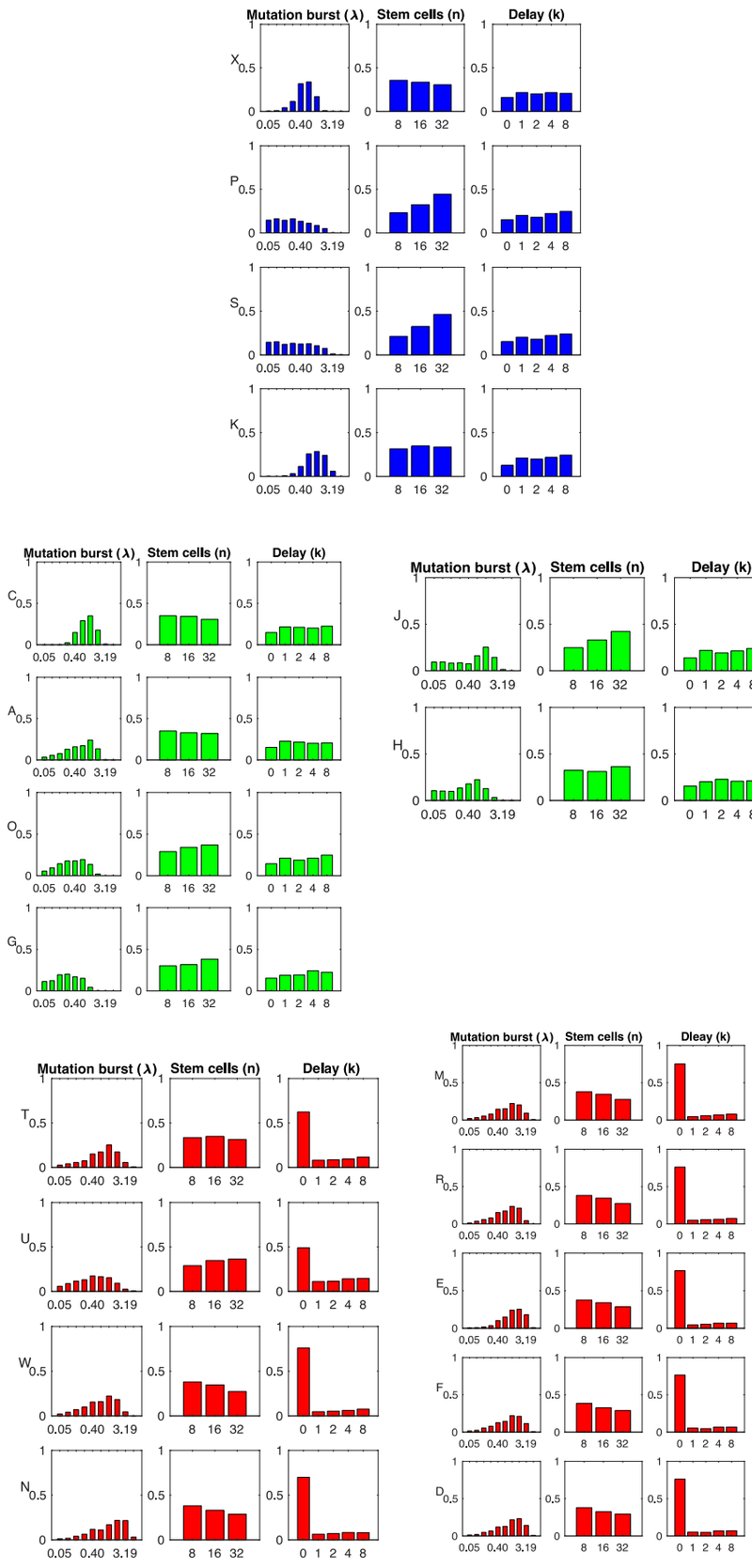


Fig. S8. Posterior parameter distributions for delayed cell mobility model. The posterior distributions for the mutation burst rate λ (log-scale), the number of stem cells *n* (log-scale), and the mobility onset delay *k* are shown for four adenomas (*A*), six nonmixing carcinomas (*B*), and nine mixing carcinomas (*C*). The letter to the left indicates the tumor name (see also Table 1), and the *y* axis shows the posterior distribution of parameters over their prior ranges.

Model M_4
Self-seeding

A - Adenomas

B - Non-mixing carcinomas

C - Mixing carcinomas

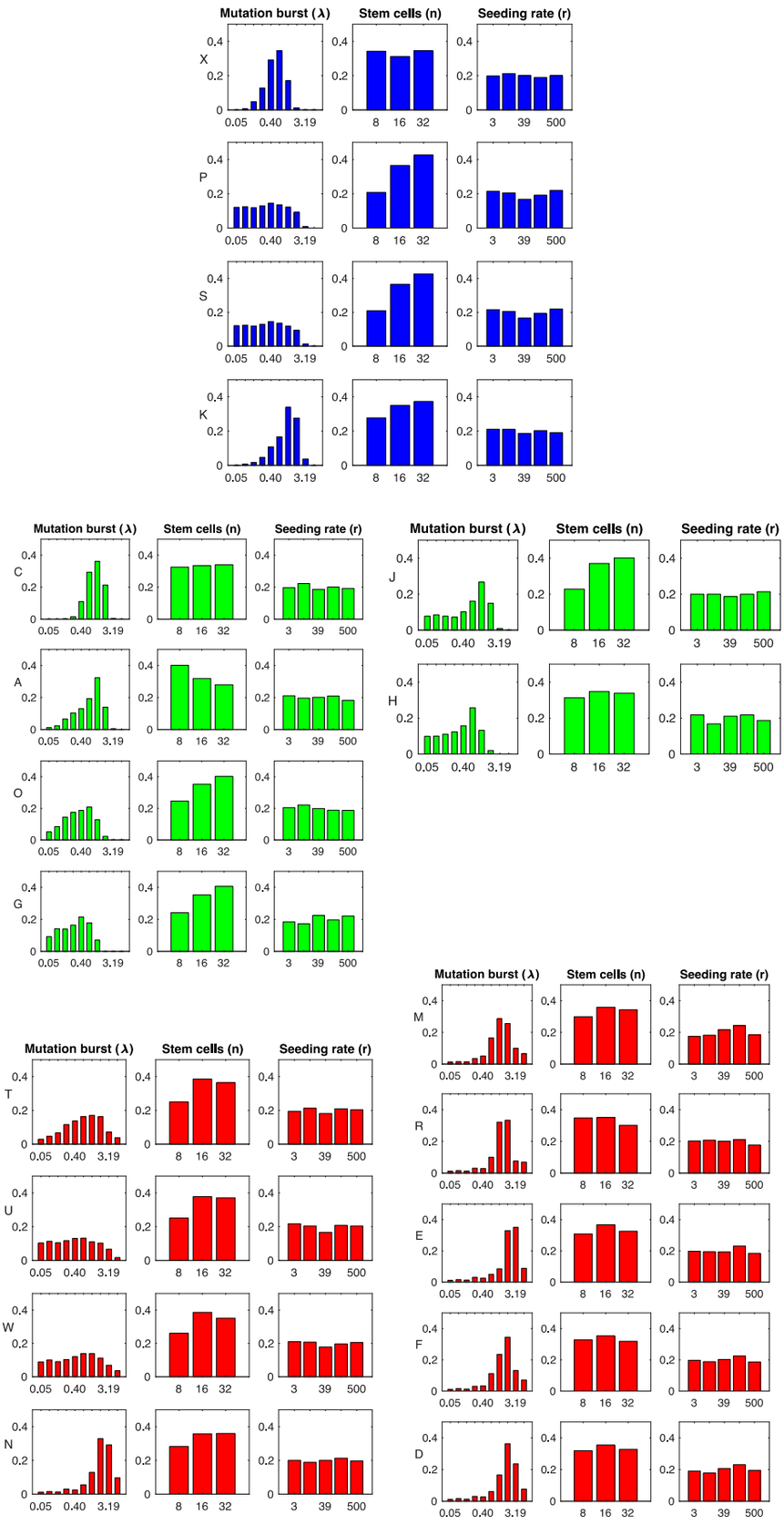


Fig. S9. Posterior parameter distributions for self-seeding model. The posterior distributions for the mutation burst rate λ (log-scale), the number of stem cells *n* (log-scale), and the self-seeding rate *r* are shown for four adenomas (*A*), six nonmixing carcinomas (*B*), and nine mixing carcinomas (*C*). The letter to the left indicates the tumor name (see also Table 1), and the *y* axis shows the posterior distribution of parameters over their prior ranges.

Other Supporting Information Files

Dataset S1 (PDF)