Modeling crypt dynamics. This part of the Supplemental Material describes in more detail the model of stem cell and non-stem cell crypt dynamics used in the simulations summarized in the article. The biological motivation for the model is given in Potten and Loeffler (1) and Potten (2) for example.

A crypt contains N stem cells after every division. The cell replication mechanism has the following form. Label the stem cells in a given generation as 1,2,...,N, and let ν_i denote the number of stem cell daughter cells born to stem cell i, $i = 1, 2, \ldots, N$. Because we assume a constant number of stem cells after each replication, the ν_i are not independent. To specify their joint distribution, we use a model that arose originally in population genetics [Karlin and McGregor (3); Cannings (4)]. Let $X_1, X_2, \ldots, X_N$ be independent random variables with distribution

$$Q(i) = \mathbb{P}(X_j = i),\ i = 0, 1, 2. \tag{1}$$

The joint distribution of $\nu_1, \ldots, \nu_N$ is then given by the joint distribution of $X_1, \ldots, X_N$, conditional on $X_1 + \cdots + X_N = N$. We assume that the same distribution applies in different cell divisions, the offspring numbers in different divisions being independent of one another.

Each of the N cells that are not designated to be stem cells after a given division initiate independent branching processes with offspring distributions that may vary with time since initiation, and all of whose progeny cells are also non-stem cells. Each such branching process goes extinct after at most a fixed number (h) of generations after initiation, because the oldest non-stem cells are lost to maintain an approximately constant total number of cells per crypt.

To simulate the evolution of a population of crypt cells replicating in this way, we have to simulate from the joint distribution of the ν_i. This can be done as follows. Denote the l-fold convolution of Q by

$$P(l, i) = \mathbb{P}(X_1 + \cdots + X_l = i),\ i = 0, 1, 2, \ldots . \tag{2}$$

The conditional distribution of ν_{j+1} given $\nu_1, \ldots, \nu_j$ is given by

$\mathbb{P}(\nu_{j+1} = i_{j+1} \mid \nu_j = i_j, \ldots, \nu_1 = i_1) =$

$$
\begin{aligned}
&= \frac{\mathbb{P}(\nu_{j+1} = i_{j+1}, \nu_j = i_j, \dots, \nu_1 = i_1)}{\mathbb{P}(\nu_j = i_j, \dots, \nu_1 = i_1)} \\
&= \frac{\mathbb{P}(X_{j+1} = i_{j+1}, X_j = i_j, \dots, X_1 = i_1 \mid \sum_{r=1}^N X_r = N)}{\mathbb{P}(X_j = i_j, \dots, X_1 = i_1 \mid \sum_{r=1}^N X_r = N)} \\
&= \frac{\mathbb{P}(X_{j+1} = i_{j+1}, X_j = i_j, \dots, X_1 = i_1)\mathbb{P}(\sum_{r=j+2}^N X_r = N - i_1 - \cdots - i_{j+1})}{\mathbb{P}(X_j = i_j, \dots, X_1 = i_1)\mathbb{P}(\sum_{r=j+1}^N X_r = N - i_1 - \cdots - i_j)} \\
&= \frac{\mathbb{P}(X_{j+1} = i_{j+1})\mathbb{P}(\sum_{r=j+2}^N X_r = N - i_1 - \cdots - i_{j+1})}{\mathbb{P}(\sum_{r=j+1}^N X_r = N - i_1 - \cdots - i_j)} \\
&= \frac{Q(i_{j+1})\, P(N-j-1, N - i_1 - \cdots - i_{j+1})}{P(N-j, N - i_1 - \cdots - i_j)}. \qquad (3)
\end{aligned}
$$

The joint law of $\nu_1, \dots, \nu_N$ is then given by

$$
\begin{aligned}
&\mathbb{P}(\nu_1 = i_1, \nu_2 = i_2, \dots, \nu_N = i_N) = \\
&\quad = \mathbb{P}(\nu_1 = i_1) \prod_{j=1}^{N-1} \mathbb{P}(\nu_{j+1} = i_{j+1} \mid \nu_j = i_j, \dots, \nu_1 = i_1) \\
&\quad = \prod_{j=0}^{N-2} \frac{Q(i_{j+1})\, P(N-j-1, N - i_1 - \cdots - i_{j+1})}{P(N-j, N - i_1 - \cdots - i_j)}. \qquad (4)
\end{aligned}
$$

Thus the numbers $\nu_1, \dots, \nu_N$ in a given generation may be simulated recursively from Eq. 4 by simulating observations from the distribution given in Eq. 3 in the order $j = 0, 1, \dots, N-2$, and noting that $\nu_N = N - \nu_1 - \cdots - \nu_{N-1}$. The convolution probabilities $P(l,i)$ can be calculated in the usual way via

$$P(l+1, i) = \sum_{r=0}^{2} Q(r) P(l, i-r),$$

with $P(1,i) \equiv Q(i)$. These can be computed once at the start of the simulation, and an observation from the distribution in Eq. 3 can then be simulated easily.

To model the evolution of the methylation patterns in the cells, we consider mutations arising at a particular CpG island as the region containing the island replicates through the cell population. We keep track of a single copy of the region in each cell for X chromosomes and two copies of the region in each cell for autosomes. The process begins with N cells containing no methylated sites in the island of interest. At each division, the CpG islands in each stem cell and non-stem cell are allowed to mutate. We chose the simplest model in which each site flips from methylated to unmethylated (or vice versa) with probability μ, independently for each site in the island.

Because we know the age of each individual in our data, we know how many divisions (g) are necessary, assuming one division per day. The method outlined above is then used to simulate the methylation patterns in a population of crypt cells (containing both stem and non-stem cells) that has evolved for g generations. We note that if non-stem cell lineages have a maximal life time of h generations, we do not need to simulate the non-stem cells until generation $g - h$. Further details of the mutation model for CpG islands appear below. A simulation run results in a population of C cells, and therefore $2C$ CpG islands (when considering an autosomal locus) or C (when considering an X-chromsome locus). From that population of islands, a random sample is taken and the statistics of the methylation patterns in the sample are recorded.

In the simulations, we used a variety of different parameter values. The mutation rate $\mu = 2 \times 10^{-5}$, and the cell division distribution in Eq. 1 is given by

$$Q(0) = Q(2) = (1 - p)/2,\ Q(1) = p,$$

where $0 \leq p \leq 1$. For the immortal cell line model, $p = 1$. Values of p for the niche model varied, as described in Table 3. The way in which the lineage initiated by a non-stem cell replicates also depends on N, as described in Table 3. Our analysis is constrained by the fact that the total number of cells in a crypt is about 2,000; in the simulations described in the article, the crypts always contain 2,048 cells. The last three divisions mentioned in Table 3 are really not divisions but reflect the observation that in a steady-state process the oldest non-stem cells no longer divide but stay around for awhile before exiting the crypt.

Time to the most recent common ancestor, MRCA, of stem cells. The distribution of the time to MRCA stem cell of a group of N stem cells evolving according to our model was found by a coalescent simulation [Kingman (5); Tsao *et al.* (6)]. First, the number of distinct ancestors, A_1, of N stem cells was found from a simulation of $\nu \equiv (\nu_1, \ldots, \nu_N)$. From an independent simulation of ν, the number A_2 of distinct parents of a random sample of size A_1 of the N was found. If $A_2 > 1$, the process was repeated, producing a series of ancestral numbers $A_1, \ldots, A_{t-1} > 1, A_t = 1$. This simulation returns a value of t for the time to the MRCA of the population of

Table 3: Parameters in the simulations.

N	p	non-stem cell divisions*
4	0.98	(0,0,1) 7; (0,1,0) 3
16	0.95	(0,0,1) 5; (0,1,0) 3
64	0.95	(0,0,1) 3; (0,1,0) 3
256	0.89	(0,0,1) 1; (0,1,0) 3
512	0.75	(0,1,0) 3

* The notation (p_0, p_1, p_2) n describes the offspring distribution, and the number n of generations it applies for. Here, p_i is the probability of a cell division producing i copies.

N stem cells. The whole process can be repeated as many times as required to estimate properties of the number of divisions to the MRCA.

We note that this framework provides quite a general methodology for examining methylation patterns in CpG islands in colon crypts. For example, the number of stem cells can be allowed to fluctuate with time (this amounts to changing the value of N in different generations), and the distribution of mutations in an island can be much more complicated than the one described here. For example, it can allow the distribution of mutations in offspring cells to depend on the current methylation pattern in the parent cell. It remains to be seen what realistic parameter values for such models should be. We note that because the number of cells in a crypt is quite small ($\approx 2,000$), there is no need to simulate the sample using coalescent methods. In any case, the presence of non-stem cells makes this approach quite difficult to implement. Computer programs used in this work are available from Simon Tavaré (stavare@hto.usc.edu).

The effects of contamination on the number of unique tags. In the article we gave a lower bound of 95% for the proportion of epithelial cells present in a sample from a given crypt. The remaining 5% of cells are nonepithelial cells. In this section, we discuss the likely effects of such contamination in the context of the number of unique tags observed among different crypts. The sample from a crypt contains approximately 1,000 cells, and therefore about 50 contaminating cells. In a sample of 8 molecules from this mixture (8 being the typical experimental sample), the number of contaminating molecules therefore has approximately a binomial distribution with success probability 0.05. The probability of no contaminating molecules in the sample is 0.66, the probability of 1 is 0.28, and the probability of 2 or more is 0.06.

We might suppose that each contaminating molecule contributes a single unique tag to the number of unique tags we observe in the data from a given crypt. Let U denote the observed number of unique tags in a sample, let V denote the number of these arising from epithelial cells, and let J denote the number coming from contaminating cells. Then $U = V + J$, where J and V are independent. J is approximately binomially distributed with variance $=$ $8 \times 0.05 \times 0.95 = 0.38$, and hence the variances σ_U^2 and σ_V^2 of U and V satisfy $\sigma_U^2 = \sigma_V^2 + 0.38$.

In Table 2 of the article, we compare the intracrypt variance of the number of unique tags as estimated from the data with those predicted under our model. The models assume no contamination, and therefore they need to be compared with the intracrypt variance to be expected in the data when contamination is removed. We see from the previous paragraph that the expected value of this intracrypt variance, σ_V^2, is given by $\sigma_U^2 - 0.38$. Thus, a rough allowance for contamination could be made by reducing the observed variance in Column 1 of Table 2 of the article, which estimates σ_U^2, by about 0.38. Comparing these reduced estimates with the model results given in Table 2 of the article, we see that our conclusions are indeed strengthened. This analysis of the possible role of contamination is somewhat simplified, but we conclude that low levels of potential contamination should not alter the thrust of our results.

1. Potten, C. S. & Loeffler, M. (1990) *Development* **110**, 1001–1020.
2. Potten, C. S. (1996) *Stem Cells* (Academic Press, New York).
3. Karlin, S. & McGregor, J. (1962) *Proc. Natl. Acad. Sci. USA* **51**, 598–602.
4. Cannings, C. (1974) *Adv. Appl. Prob.* **6**, 260–290.
5. Kingman, J. F. C. (1982) *J. Appl. Prob.* **19A**, 27–43.
6. Tsao, J., Yatabe, Y., Salovaara, R., Järvinen, H. J., Mecklin, J.-P., Aaltonen, L. A., Tavaré, S. & Shibata, D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 1236–1241.