Supplementary Materials and Methods:

Single molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization

Andrea Sottoriva[1,2,3], Inmaculada Spiteri[2], Darryl Shibata[4], Christina Curtis[3,†], Simon Tavaré[1,2,5,†]

[1] Department of Oncology, University of Cambridge, Cambridge CB2 2XZ, UK

[2] Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK

[3] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033

[4] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033

[5] Department of Biological Sciences, University of Southern California, Los Angeles, 90089

[†] Correspondence and requests for materials should be addressed to:

Christina Curtis (ccurtis@usc.edu) or

Simon Tavaré (simon.tavare@cancer.org.uk)

The tumor growth model

Colorectal cancer is organized into glandular structures of about 8,000-10,000 cells each (1). Proliferation occurs within the gland and when a certain number of cells are generated, the gland splits into two following a process of gland fission. Thus, we can think of the glands as entities that grow and divide when a certain volume is reached. The exact time when the gland divides and the molecular patterns that it contains depend on the characteristics of the gland and its parameterization, such as the cancer stem cell fraction, methylation rate, etc. At first we are only interested in calculating the phylogeny of the glands but we neglect their content and the corresponding time scales. We leave the calculation of the intra-gland details for later. This first part of the model is called *crecspace* and it is a cellular automaton model in which the space is represented as a 3-dimensional lattice of points. Each point in the lattice is the size of a gland (8,000-10,000 cells) and can be a cancer gland or normal tissue. The tumor is initialized with a single cancer gland in the center of the lattice and the rest occupied by normal tissue (not modeled). The cancer grows and invades the surrounding tissue, as a result of a selective growth advantage. At each time step, we simulate gland division by picking a gland at random and choosing a neighboring lattice site where the new gland will be put. To make space for newly generated glands, existing glands are shifted outward, emulating the proliferative pressure that the tumor imposes on its surroundings. At the same time, we keep track of the phylogenetic history of the glands from the first founding cell to tumor resection. For each node in the phylogenetic tree we save the position of the gland in the tumor and at the end of the simulation the tree is saved in Newick

format. We note that at this point the glands are still empty objects for which the parental history and spatial position within the tumor are known, but the molecular profile is ignored. The model is summarized by the following pseudocode:

while volume < 16 million glands (7.5 cm tumor)

do

choose random gland

choose random neighbor

shift outwards all existing glands along the direction of the neighbor

position the newborn gland into the resulting void

create a branch in the phylogeny and save gland positions

end

This tumor growth algorithm is fully spatial and three-dimensional and achieves a striking efficiency, simulating 16 million glands and their phylogeny within a $200cm^3$ space in about 10 minutes. Moreover, if needed, this model can be easily extended to capture further complexity, for example by introducing different gland types or microenvironmental interactions.

The second part of the model is called *crecmeth* and takes as input the phylogenetic tree of the cancer glands and computes the within-gland dynamics using a population genetics approach that neglects the spatial information of the single cells in the gland. The loaded tree contains the information of the whole tumor, yet the data come from only a small fraction of the whole tumor mass. Hence, to achieve further speedup, depending on the

sampling scheme the tree is pruned and only the sub-tree that contains the history of the sampled glands is kept. For instance if we had sampled 2 glands in position (x_1,y_1,z_1) and (x_2,y_2,z_2) of the malignancy, we pick these two glands from the tree, save the related phylogeny and discard the rest. It is critical that this virtual sampling step emulate exactly what happened in reality: if we sampled one area of the tumor, we similarly sample from the same area in the virtual tumor. This pruning step dramatically reduces the complexity of the phylogeny and avoids simulating billions of cells that would have never been sampled in reality. The resulting tree is small, with tens or at most hundreds of leaves, depending on the number of samples collected from the actual patient sample.

In *crecmeth* each gland contains approximately 8,000 cells with different patterns or methylation haplotypes. Every haplotype contains a counter of the cells within the gland that have that particular haplotype. Cells can also be of different types if a cancer stem cell organization is present: cancer stem cells (CSC), transit amplifying cells (TAC) and differentiated cells (DC). At each cell division, cancer stem cells generate a new CSC with probability ψ and a TAC with probability $1-\psi$. The TAC divides G times before differentiating into a DC. If a classical, flat model of malignant growth is simulated, all cells in the gland are assumed to be stem-like cells and have unlimited replicative potential with $\psi=1$.

The simulation is initiated with the gland at the root of the phylogenetic tree containing one single cell with an initial haplotype. The cells in the gland divide and the haplotype counter is increased together with the counter of the gland population. Mutations are of two types: methylation occurring at rate μ_m

and demethylation occurring at rate μ_d. Whereas in principle μ_m and μ_d are equal, they may diverge due to the context-dependent methylation factor χ discussed later on. Assuming a binary string, *H,* corresponding to the methylation pattern of a cell with N_0 unmethylated CpGs and N_1 methylated ones, the probability that after a division the pattern changes is:

$$P(X+Y>0)=1-(1-\mu_m)^{N_0}\left(1-\mu_d\right)^{N_1}$$

where X denotes the number of newly methylated CpGs and Y the number of newly unmethylated CpGs. In general, the distribution of X and Y is given by:

$$P\left(X=l,Y=n\right)=\left(\begin{array}{c} N_0 \\ l \end{array}\right)\mu_m^l(1-\mu_m)^{N_0-l}\left(\begin{array}{c} N_1 \\ n \end{array}\right)\mu_d^n(1-\mu_d)^{N_1-n}$$

Hence, if a mutation occurs, the probability of having *l* methylation events and *n* demethylation events with *l*+*n* > 0 is:

$$P(X=l,Y=n\,|\,X+Y>0)=\frac{P(X=l,Y=n)}{P(X+Y>0)}$$

Since methylation errors are rare per cell division ($\mu<<1$), we consider only the events where $X+Y \leq 2$. It is also believed that the actual methylation status of a CpG island influences the methylation rate in such a way that methylated regions in turn have a higher methylation rate. This mechanism, referred to as context-dependent methylation, acts as a sort of positive

feedback that tends to favor methylation over demethylation on already methylated CpG islands. Whether this process happens in reality and with what magnitude is still unknown, however Nicolas and co-authors (2) reported a substantial increase in the quality of the inference when context-dependent methylation is considered. For this reason, we introduce a new parameter χ that changes the methylation rate μ_m depending on the methylated fraction π of the whole pattern, as the following equation illustrates:

$$\mu_{m'} = \mu_m + \mu_m \cdot \pi \cdot \chi$$

with $\mu_{m'}$ always <1. When a mutation occurs a new haplotype is allocated and the corresponding mutated cells are relocated from the old to the new pattern. Computing the population dynamics of the gland in this manner is a very straightforward task since it is a matter of sampling events from binomial distributions. When the total number of the cells in the gland becomes larger than 8,000 the gland splits into two by fission, and half of the cells are allocated to each daughter gland in a random fashion. Other forms of gland fission are possible, such as a single founding cell for the new gland. Although we do not consider this option in our model, we predict that it will further enhance the level of heterogeneity (**Figure S7**). Once the gland division is completed, the phylogenetic tree is descended. Assuming no recombination, the two alleles of a cell mutate independently. Moreover, in the samples the alleles are mixed together so the information of which two alleles belonged to the same cell is lost. When the end of the phylogenetic tree is reached, the gland content is saved in a standard ASCII format with methylation patterns

as binary strings. The architecture of *crecmeth* is summarized by the following pseudocode:

load phylogeny

sample tumor as done in the data

prune the tree not sampled

for all tree nodes

while gland size < 8,000

for all haplotypes in the gland do

S ← number of stem cells

new stem cells $S_n \sim Bin(\psi, S)$

new first stage TA cells $T_{n1} \leftarrow S - S_n$

mutated stem cells $S_m \sim Bin(P(X+Y)>0, S+S_n)$

mutated TA cells $T_{m1} \sim Bin(P(X+Y)>0, T_{m1})$

handle mutation events with $P(X = k_i, Y = k_j \mid X + Y > 0)$

for all TA stages do

simulate division and mutation

last state is DC, no further division occur

end

allocate all new haplotypes for mutated cells

simulate apoptosis as random cell death

update total cell population

end

end

end

end

In *crecmeth* the within-gland spatial information is neglected to make the problem computationally tractable. We feel that this approximation does not introduce a relevant bias in our analysis, since spatial information at a cellular level is currently not recorded during laser capture microdissection of tumor specimens and subsequent high-throughput sequencing. We note that simulating spatial features of single cells based on observed data from a mass of 10^{11} cells is not informative, as we cannot observe dynamics at that level of detail. While spatial information is important, it need not be handled across 5 orders of magnitude (from microns to decimeters), which is currently impractical from both an experimental and computational perspective. Despite such methodological limitations, our approach records spatial information across 3 orders of magnitude, from 100 μm glands to 10 cm wide neoplasms.

In our study we resected a volume of 0.5 cm^3 on the left (L) and on the right (R) side of colorectal cancers, and microdissected 4-5 glands per side. After simulating the phylogenesis with *crecspace* we save only the glands within these two areas of the tumor and their corresponding phylogenetic tree. This further reduces the complexity of the tree. When *crecmeth* loads the tree, it virtually samples 4-5 glands from each of these two areas and computes the corresponding molecular patterns. The simulation contains more than 16 million glands and represents a tumor that is 7.5 cm in diameter. This modular structure for the model is highly adaptable and the level of detail can be adjusted to suit the sample size and precision.

The statistical inference method

The inference scheme reported in the **Materials and Methods** in

essence dictates that we sample a set of parameters from a certain prior distribution, run the model with those parameters, and accept only those simulations for which the output is close to the observed data. This algorithm can be easily parallelized by computing all the simulations first, which are independent, and applying the rejection step at the end. Therefore, we sample millions of parameter sets $\theta=(\xi,\psi,G,\mu,\chi,a)$ from uniform prior distributions and we run the model with those parameters, generating as output methylation pattern distributions. Not every combination of parameters is acceptable, in the sense that some are not able to sustain tumor growth (e.g. very low CSC fraction together with high apoptosis may kill off the whole tumor population). For this reason a small subset of the sampled parameters will be rejected early on. The distribution of the parameters that are able to generate malignancies is reported in **Figure S8**. We aimed to achieve uniform distributions for those patterns with a particular focus on the CSC fraction, which is of great interest. To do so the symmetric division rate parameter cannot be uniform, as can be appreciated from **Figure S3**. When considering all the possible scenarios, the tumor progression time from the first tumorigenic cell to a large carcinoma (tumor age) is reported to vary between 12 and 39 months. For each simulation we calculate the summary statistics of the methylation pattern output using the measures (S_p, S_d, S_s, S_w, S_k, S_h) reported in the main manuscript. The resultant file contains the set of parameters $\theta=(\xi,\psi,G,\mu,\chi,a)$ followed by the set of summary statistics for z glands as in $(S^1p,S^1d,S^1s,S^1w,S^1k,S^1h)$, …, $(S^zp,S^zd,S^zs,S^zw,S^zk,S^zh)$. To allow comparison between different summary statistics, each of them is normalized to mean=0 and SD=1. The overall distance between a pair of

glands *i* and *j* is:

$$S^{i,j} = \left|S_p^i - S_p^j\right| + \left|S_d^i - S_d^j\right| + \left|S_s^i - S_s^j\right| + \left|S_w^i - S_w^j\right| + \left|S_k^i - S_k^j\right| + \left|S_h^i - S_h^j\right|$$

Thus we measure the distance between a real tumor U and a simulated tumor V containing z glands each by pairing them gland by gland, maintaining gland information (e.g. left glands with left glands and same number of reads) as follows:

$$\rho = \frac{1}{6z}\sum_{i=1}^{z} S^{U_i,V_i}$$

As in the standard ABC scheme, we impose a threshold ε and accept all simulated particles that yield ρ() < ε.

Despite the approximation, ABC is extremely efficient and permits us to perform inference with a slight sacrifice in accuracy but enhanced speed in comparison to other Bayesian approaches. When inference is performed on complex systems, ABC is often the only feasible solution. Using *SCAI*, we interrogated the following parameters of colorectal cancer dynamics, assuming uniform priors:

- The CSC fraction $\xi \in [0,1]$
- The number of TACs $G \in \{0,1,2,3,4,5\}$ (for the CSC model only, $\xi < 1$)
- The methylation/demethylation rate $\mu \in \{10^{-6},10^{-5},10^{-4},10^{-3}\}$
- The context-dependent methylation factor $\chi \in \{1,2,5,10\}$ times the

normal methylation rate

- The apoptosis rate $\boldsymbol{a} \in \{0, 0.01, 0.1\}$ (fraction of dead cells per cell division)

The parameter ranges are chosen to probe a wide and realistic set of values, with limitations due to the computational feasibility. For instance, with **μ** reported on the order of 10^{-5} from literature, we argue that 4 orders of magnitude well describe the range of possible methylation rates, with the values outside such a range being unlikely.

The combination of ξ and *G* determines the symmetric division rate $\psi \in [0, 1]$, which is not uniform since values of $G > 0$ generate a large population of transit amplifying cells and consequently a lower CSC fraction. Calculating ξ given ψ and *G* is not straightforward and requires the pre-computation of a large table from which to extract ψ once ξ and *G* have been sampled from their uniform distributions. The relationship between the three parameters is illustrated in **Figure S3**; different curves correspond to different values of *G*, where a higher CSC fraction corresponds to a more clonal growth model and more uniformly tumorigenic cells. The CSC fraction is an important parameter, and we note that the tumor age γ is the result of the CSC fraction (which determines the clonogencity of the tumor) and the cell cycle time. Finally, the distribution of parameters of all simulations that survive the rejection step represent the posterior distribution of that parameter given the observed molecular data.

The Spatial Cell Ancestral Inference framework

Methylation patterns capture the processes involved in tumor phylogeny, but some mechanisms may not leave a detectable signature in the methylation patterns. Additionally, the summary statistics may not fully capture the information contained in the methylation patterns. To test our *SCAI* approach we performed inference on synthetic data generated with our model to verify which parameters we can recover correctly and which ones we fail to estimate. We generated a synthetic dataset containing 9 tumors each 7.5 cm in diameter (130 billion cells) with 4 glands resected per tumor side and 1,000 reads per gland for a 16 CpG-long locus. We selected different parameter sets to verify the power and limitations of the inference for different values and scales of the parameters. **Figure S2** shows the posterior distributions of the parameters for the synthetic dataset. The parameters that determine tumor organization, such as ξ and ψ, are all recovered with a considerable level of precision. The ability to distinguish the CSC fraction from the molecular clock signature is very important in our analysis since we want to address in depth the major question of cancer stem cells in human malignancies. The short life of TACs does not allow a strong signature of their structure to be detected in the methylation patterns. Indeed, methylation events that occur in TACs are very rare because of the limited number of cell divisions they undergo before becoming differentiated (DCC). We note that it is extremely hard to infer the TAC signal from the molecular data using this approach. Nonetheless, our inability to correctly recovery this parameter does not significantly impact the inference on the CSC fraction, which is consistently recovered correctly (**Figure S2,** column **A**).

The parameters that determine the mutation rates, such as μ and χ, are also recovered with precision. These values define the speed of the molecular clock and are crucial to the study of methylation errors in tumors since the methylation rate has been only been crudely measured in normal tissue, with a value on the order of $2x10^{-5}$ (3). The ability to estimate mutation parameters with high precision is fundamental to understanding the process of mutation and progression in cancer, but when we use mutations as molecular clocks it is also the key to inferring many other parameters that directly or indirectly depend on those rates. We report that tumor age is hard to infer for small cancer stem cell fractions that yield slow tumor growth. We find that this occurs because tumors with low ξ have variable ages, ranging from a few months to a few years for extremely small CSC populations (ξ<0.1%). Conversely, the large majority of tumors with high CSC content have a very small range of ages, between 8 and 10 months. Thus it is possible to identify clear peaks in tumor age for higher values of ξ, whereas for low values the posterior distribution is spread across a broad time-span. Despite a few limitations, the *SCAI* framework accurately recovers many important parameters that cannot be measured *in vivo*. The same results are derived if we consider a synthetic data with 8 CpG-long molecular clocks (**Figure S9**).

Moreover, although copy number alterations of the IRX2 locus are extremely rare in CRC (4), we demonstrate that an eventual loss of the locus (simulated as a haploid molecular clock) did not affect the results (**Figure S10**). This is expected since the two alleles are in principle independent molecular clocks that are randomly collected from the tumor gland sample. Additionally, we verified that increasing the throughput per gland from 24

reads/gland to >1000 reads/gland yields significant differences in the summary statistics, as demonstrated by a simulation in **Figure S11**.

Confirmation of the results by additional molecular clocks

A key question when exploiting molecular clocks is whether different neutral loci carry the same information about tumor dynamics. Due to the stochastic nature of methylation errors, two distinct areas of the genome may exhibit different methylation patterns, but should be equivalent in terms of the information they convey. This question has remained unresolved as past studies showed that different clocks reveal different behaviors (1). The fundamental differences between molecular clocks are their length and their methylation rate that determine the clock precision. A fast clock records more mitotic events, however if it is also short (a few CpGs) it is at risk of becoming fully methylated and so losing the relationship with cell division. This process of *saturation* should be avoided. The accuracy of the inference is also crucially dependent on the number of reads per sample.

To validate our findings for the IRX2 locus, we interrogated methylation haplotype data based on 2 further molecular clocks for tumor Z: ZNF454 (200bp) on chromosome 5 (16 CpGs, >1,500 reads/gland) and SLC5A7 (111bp) on chromosome 2 (6 CpGs, 300 reads/gland) for which we verified neutrality (**Figure S4**) and report no case of copy number alteration within the TCGA cohort (4). Primer sequences and PCR conditions for the loci IRX2 (201 bp), ZNF454 (200 bp) and SLC5A7 (111 bp) are provided in **Table S1**. **Figure S5A** confirms that for all the clocks the glands cluster by side, indicating the same clonal organization previously discussed. We then applied the SCAI framework to those data, and observe that all three loci produced

comparable results, with IRX2 and ZNF454 in particular generating very similar posterior distributions (**Figure S5B** and **S6**). SLC5A7 showed some incongruences, likely due to its length (6 CpGs) and the lower PCR efficiency resulting in 5 times fewer reads. In light of these results we underline the importance of using high-throughput data for inference-based analyses with an average depth of >1,000 reads per sample and neutral genomic loci containing at least 8 CpGs.

Gland clustering and intra-gland phylogeny

Clustering of different glands from the same tumor (**Figure 2B** and **S5A**) was performed by hierarchical clustering using a measure of mitotic distance between two sets of patterns as a measure of similarity. To calculate such measure, we first paired the patterns with the minimal mitotic distance (calculated as Hamming distance) from the two sets, starting from the closest pair. The sum of all distances of the pairs was considered as the similarity measure between two patterns sets (i.e. two glands). If one set had a different number of reads, the additional reads from the larger set were randomly selected and discarded.

Reconstruction of the phylogeny of the methylation patterns within the glands (**Figure 2A**) was done using a simple Neighbor-Joining (5) method based on the Hamming distance between the single methylation patterns.

References

1. Siegmund KD, Marjoram P, Woo YJ, Tavaré S, Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc Natl Acad Sci U S A 2009; 106:4828–33.

2. Nicolas P, Kim KM, Shibata D, Tavaré S. The stem cell population of the human colon crypt: analysis via methylation patterns. PLoS Comput Biol 2007; 3:e28.

3. Yatabe Y, Tavaré S, Shibata D. Investigating stem cells in human colon by using methylation patterns. Proc Natl Acad Sci U S A 2001; 98:10839–44.

4. TCGA Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487:330–7.

5. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 1987; 4:406 –425.

Tables

Locus	Sequence
IRX2	GTATATTTTGTTAGGATTGGAGTTGCGGGGATTTTTTAGTTT TTTACGGATGTTTTTTTGTACGTGTTTTTTAGGAAAATATTTA GGTTTCGGAAGCGTTTTTGAGAGTTTAGAAATGTTGGTTGA GCGTGAAATTTAGTTTAATTTATTTATAGAGTATTAGATTTAG AAACGACGTTTGAAAGAGATGTGGGTTTTATG
ZNF454	ATTAGGTTTTTGTTTTGTATTAAGGCGTAGGAAGATGTAGAT TAGCGAAAAGTTGAAGATGGGGTAAGGGAGGTTAGTTTAGG GGAGCGACGTTCGATCGTTGACGAGAGAGGATCGGGAGGG TTCGGAAGACGGAGTGGGAGGCGATAGGGCGCGTTTTTGA TTGGGATCGCGTCGAAAAGGTTTGGAAGGTAATAGAA
SLC5A7	TGTAAGAGGTTATAAAGTTTGGGCGTAGGAAATGGGTAGAG GGGGTCGAGGAAGGGTCGTAGGGGGTCGGGAGAGTATCG GGTTGTTGCGGAGAGGAATTTGTTTGTTTGG

Table S1. Genomic sequences of molecular clocks.

Tumor	Single glands	Left side	Right side	All tumor
CT	2.54	3.01	3.09	3.23
CU	2.19	2.50	3.32	3.48
CX	1.74	2.42	1.23	2.20
HA	2.24	2.69	2.57	2.77
Z	1.43	1.57	2.36	2.50

Table S2. Summary of intra-tumor heterogeneity. The Shannon Index (SI) was used to summarize the dataset in terms of intra-tumor heterogeneity at

different spatial scales: intra-gland heterogeneity (average of SI across all glands), index of the left side, index of the right side, and overall index of the whole tumor. The data show a clear trend of heterogeneity directly proportional to spatial distance.

Supplementary Figures

Figure S1. Validity of the IRX2 molecular clock. Our analysis is based on the neutrality of methylation mutations at specific genomic loci. Data from human autopsy-derived normal samples show that the IRX2 locus exhibits age-related increase in methylation in mitotic tissue (colon), but consistently low methylation in non-dividing tissues (heart and brain) and neutrophils. This is evidence of the neutral behavior of the locus.

Figure S2. Validation of the SCAI framework with synthetic data. With our tumor growth model we generated synthetic data and verified that for different parameter combinations our framework is indeed able to infer the correct values (denoted by red triangles). We report a very high success rate for all the parameters except for the number of transit amplifying stages, G. The reason for this is that short-living TACs leave a very dim molecular signal in the methylation data. Yet, this problem does not affect our ability to correctly recover the CSC fraction (column **A**).

Figure S3. Tumor organization parameters. The CSC fraction ξ depends on the symmetric division rate ψ and the number of transit stages G. When ξ is

large the tumor tends to exhibit a classical or purely clonal organization, whereas low values correspond to a CSC organization. Large values of *G* also decrease the CSC fraction by amplifying the differentiating compartment of the tumor population.

Figure S4. Validation of ZNF454 and SLC5A7 as molecular clocks. Both ZNF454 **(A)** and SLC5A7 **(B)** display consistently low methylation levels in non-dividing tissues (heart and brain) and neutrophils, but age-related methylation in mitotic tissues like the colon. All samples refer to normal human tissues derived from autopsies.

Figure S5. Inference on tumor Z using different molecular clocks. For tumor Z, two additional molecular clocks were assayed: ZNF454 and SLC5A7. For all loci, glands cluster by tumor side **(A)**, moreover they all yield very similar results in terms of the posterior distributions of tumor parameters **(B)**. Relatively small differences in results were noted for the SLC5A7 locus, we argue that that is due to the much lower yield in terms of reads per gland (5 fold lower) than the others clocks, as well as due to the reduced length of the clock (6 CpGs). Overall, these results demonstrate that for the same tumor, different molecular clocks carry comparable information on the mitotic history of the malignancy.

Figure S6. Inferred parameters *G* and *a* for ZNF454 and SLC5A7. For the transit amplifying stages parameter *G* and the apoptosis rate ***a,*** we report agreement of multiple molecular clocks.

Figure S7. Heterogeneity induced by different gland splitting algorithms. In our model we implemented the gland fission mechanism as the equal division of the cell population into the two daughter cells. Although this is the most commonly advocated method of gland splitting, an alternative scenario is to consider a single founding cell for the new gland. We predict that this would substantially increase the level of heterogeneity in the glands, due to the increased number of cell divisions required by the founding CSC to grow the new gland. We do not consider this alternative process in our analysis.

Figure S8. Prior distributions used in SCAI. We used non-informative priors for our analysis. However, some parameter values are less likely to yield a tumor. For example, a high apoptotic rate may completely extinguish the CSC population when this is small enough. For this reason, the symmetric division rate and apoptosis rate are not taken to be uniform. The tumor age is a secondary parameter that results from the combination of all the other parameters.

Figure S9. Validation of the SCAI framework with an 8 CpG-long synthetic dataset. Using synthetic dataset with a molecular clock made by 8 CpGs produces very similar results to a synthetic dataset with 16 CpGs-long clocks.

Figure S10. Patient-specific inference results for a single copy of IRX2. Inference results considering a haploid version of the IRX2 locus in each cell

(e.g. due to copy number loss), are very similar to the results obtained by considering 2 copies of IRX2 per cell. This confirms the robustness of our analysis to copy number alterations. Moreover copy number changes of the IRX2 locus in CRC are reported in an extremely limited number of cases (~0.3%).

Figure S11. Expected variation in summary statistics for different sequencing depths. The simulation of different sampling depths from the same gland reveals a considerable difference in summary statistics, particularly for the number of patterns, singletons, Kolmogorov distance and Shannon index.