# Genetic reconstruction of individual colorectal tumor histories

**Jen-Lan Tsao[†], Yasushi Yatabe[†], Reijo Salovaara[‡§], Heikki J. Järvinen[¶], Jukka-Pekka Mecklin[‖], Lauri A. Aaltonen[§], Simon Tavaré[††], and Darryl Shibata[†,‡‡]**

[†]Department of Pathology, University of Southern California School of Medicine, Los Angeles, CA 90033; Departments of [‡]Pathology and [§]Medical Genetics, Haartman Institute, FIN-00014, University of Helsinki, Finland; [¶]Second Department of Surgery, Helsinki University Central Hospital, FIN-00029, Helsinki, Finland; [‖]Jyvaskyla Central Hospital, FIN-40620, Jyvaskyla, Finland; and [††]Departments of Biological Sciences, Mathematics, and Preventive Medicine, University of Southern California, Los Angeles, CA 90089

Communicated by Harry Rubin, University of California, Berkeley, CA, December 2, 1999 (received for review September 10, 1999)

**It is difficult to observe human tumor progression as precursor lesions are systematically removed. Alternatives to direct observations, commonly used to reveal the hidden past of species and populations, are sequence comparisons or molecular clocks. Noncoding microsatellite (MS) loci were employed as molecular tumor clocks in 13 human mutator phenotype (MSI$^+$) colorectal tumors. Quantitative analysis revealed that specific patterns of somatic MS mutations accumulate with division after loss of mismatch repair (MMR). Tumors had unique patterns of MS mutation, and, therefore, based on this model, each tumor had its own unique history. Loss of MMR occurred very early relative to terminal clonal expansion, with an estimated average of 2,300 divisions since loss of MMR and 280 divisions since expansion. Contrary to the classical adenoma-cancer sequence, MSI$^+$ adenomas were nearly as old as cancers (2,000 versus 2,400 divisions since loss of MMR). Negative clinical examinations preceded six tumors, independently documenting an absence of visible precursors during early MSI$^+$ adenoma or cancer progression. These findings further extend a window beyond visible progression since loss of MMR appears to start a genetic phase involving clone sizes or phenotypes below a threshold of clinical detection. This previously occult prologue before visible neoplasia is longer and therefore likely more important than generally appreciated.**

The mutations present in colorectal tumors are diverse, suggesting a variety of pathways to cancer (1–3). Which of the many potential pathways do individual tumors follow to cancer? Tumor histories based on direct observations are problematic. Adenomas are routinely removed, and surveillance intervals necessary to observe the entire progression to cancer may span decades. Adenomas can persist for years, with years before the appearance of cancers (4–6).

The precursors of mutator phenotype (MSI$^+$) cancers have been extremely difficult to study because adenomas are not markedly increased in hereditary nonpolyposis colorectal cancer patients (3, 7). MSI$^+$ colorectal cancers are deficient in DNA mismatch repair (MMR) and have greatly elevated mutation rates, especially at microsatellite (MS) loci (3). MSI$^+$ tumors appear to progress more rapidly than repair-proficient cancer (3, 7). Of note, mice and rare humans with inherited MMR deficiencies are tumor prone but otherwise phenotypically normal and accumulate somatic MS mutations in histologically normal cells (8–11). Therefore, it is possible that at least some of the mutations present in MSI$^+$ tumors accumulate before visible neoplasia.

To overcome limitations imposed by direct observations, we hypothesize that the unique histories of individual tumors are recorded by their somatic mutations. Progression is thought to occur through successions of selection and clonal expansion (1–3). The final tumor eventually arises from a single cell that represents the last bottleneck, regardless of the number or sizes of prior waves of clonal expansion. The rest of the lineages are dead ends. Therefore, mutations common to all tumor cells accumulate along the single lineage preceding this final founder cell whereas heterogeneous mutations may arise with clonal expansion (Fig. 1).

We used a quantitative approach to infer MSI$^+$ tumor histories based on the analysis of the common and unique somatic MS mutations that may accumulate along such a progression pathway. The start of this pathway is marked by the somatic loss of MMR, which increases mutation rates in noncoding CA-repeat MS loci ≈100-fold (12, 13) and allows them to function as molecular tumor clocks (14–16). Therefore the majority of MS mutations reflect divisions after loss of MMR. Note that changes in MS allele lengths are thought to arise with slippage during DNA replication (17, 18), so mutations are coupled with cell division. Differences between the final founder cell and its germline genotype reflect mitotic divisions that occur along a single lineage preceding the founder cell. The longer the intervals between loss of MMR and final tumor formation, the greater may be the differences between germline and tumor genotypes. The genotype of the final founder cell can be inferred by determining the most common allele present among the tumor cells. Polymorphic tumor alleles, which can only arise after clonal expansion, provide information on numbers of divisions after the final bottleneck.

## Materials and Methods

**Specimens.** DNA samples were extracted from formalin-fixed microscopic tissue sections (14) of 13 tumors from nine male patients. In most cases, tumors were subdivided into multiple regions based on phenotype or topography. Clinical information was obtain from medical records. Surveillance intervals were defined by times of colonoscopy or surgery. Patients with hereditary nonpolyposis colorectal cancer had germline mutations (Patients I, IV, and VI–IX, hMLH1; Patient III, hMSH2) confirmed by sequencing. The histories of Patients II and V did not meet criteria for hereditary nonpolyposis colorectal cancer. The MSI$^+$ colorectal cell line HCT 116 [hMLH1-deficient (19)] was obtained from American Type Culture Collection and had a mitotic rate of approximately one division per day.

**MS Analysis.** To simplify analysis, X chromosome CA-dinucleotide repeat MS loci and male patients were used. Tumor DNA was diluted before PCR such that ≈20–80% of assays produced products. Every measurement of a MS locus therefore essentially represents a single cell because MSI$^+$ tumors characteristically lack aneuploidy (20). Products were labeled with $^{33}$P-dCTP (NEN) incorporated during 38–43 PCR cycles and were analyzed on 6% denaturing polyacrylamide sequencing gels

Abbreviations: MS, microsatellite; MMR, mismatch repair.

[‡‡]To whom reprint requests should be addressed at: Department of Pathology, 1200 North State Street, Box 736, University of Southern California School of Medicine, Los Angeles, CA 90033. E-mail: dshibata@hsc.usc.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

**Table 1. Tumor ages**

| Patient | Tumor | *N** | $S^2_{alleles}$ | Expansion divisions | $S^2_{loci}$ | Total divisions | Age, years | 95% CI, years | Clinical interval |
|---|---|---|---|---|---|---|---|---|---|
| I | Adenoma/Cancer-1 | | | | | | | | |
| | Adenoma 1.0 cm | 30 | 1.6 | 350 | 8.3 | 1,900 | 5.2 | 2.3–7.2 | — |
| | Cancer Dukes' C | 29 | 0.88 | 190 | 7.7 | 1,700 | 4.6 | 2.0–6.6 | — |
| | Cancer-2 Dukes' B | 30 | 1.6 | 350 | 9.2 | 2,100 | 5.7 | 2.5–7.9 | 0.5 yr |
| II | Adenoma/Cancer | | | | | | | | |
| | Adenoma 1.0 cm | 28 | 1.3 | 280 | 15.8 | 3,400 | 9.2 | 3.6–13 | — |
| | Cancer Dukes' B | 28 | 1.2 | 260 | 9.8 | 2,100 | 5.9 | 2.4–8.3 | — |
| III | Adenoma-1 0.5 cm | 25 | 0.96 | 210 | 7.5 | 1,700 | 4.6 | 1.9–6.6 | 2.0 yr |
| | Cancer Dukes' B | 29 | 0.42 | 91 | 10.3 | 2,200 | 5.8 | 2.5–8.2 | — |
| | Adenoma-2 0.5 cm | 27 | 1.9 | 420 | 6.7 | 1,600 | 4.4 | 2.0–6.3 | 2.3 yr |
| IV | Cancer Dukes' D | 21 | 3.0 | 660 | 9.0 | 2,200 | 6.1 | 2.2–8.6 | — |
| V | Cancer Dukes' D | 28 | 1.3 | 280 | 6.4 | 1,500 | 4.1 | 1.6–5.7 | — |
| VI | Adenoma 1.1 cm | 23 | 0.92 | 200 | 5.7 | 1,300 | 3.6 | 1.3–5.2 | 3.8 yr |
| | Cancer Dukes' A | 23 | 0.46 | 100 | 9.0 | 1,900 | 5.1 | 2.0–7.6 | 3.8 yr |
| VII | Cancer Dukes' B | 26 | 1.3 | 280 | 15.0 | 3,200 | 8.8 | 4.1–13 | — |
| VIII | Cancer Dukes' B | 24 | 0.87 | 190 | 10.7 | 2,300 | 6.3 | 2.7–9.2 | — |
| IX | Cancer Dukes' B | 24 | 1.6 | 350 | 22.3 | 4,700 | 12.9 | 5.3–19 | 1.0 yr |
| | Average | | 1.3 | 280 | 10.2 | 2,300 | 6.2 | | |

*Number of MS loci examined.

and a PhosphorImager (Molecular Dynamics). Lengths different from germline were considered to be from the tumor. When distributions included the germline length, 40–50% of genomes were considered to originate from contaminating normal cells (estimated by visual inspection for each tumor), and frequencies at the germline size were reduced by this amount to approximate the tumor MS length distribution.

Defining an $MSI^+$ tumor genotype can be problematic because their cells rapidly acquire further mutations. Multiple measurements at each locus are necessary to characterize a polymorphic tumor population. PCR after dilution of tumor DNA typically yields a variety of different alleles distributed around a single mode. Estimation of a mode from a tumor repeat length distribution is experimentally simpler than estimation of its mean because germline alleles from contaminating normal cells can be more easily eliminated. We summarize each MS locus by the difference ($\Delta_{germline}$) between the germline length and the mode of the tumor repeat length (Fig. 1*b*). At least 10 molecules were amplified at each locus until its mode became evident. We also estimate the sample variance ($S^2_{alleles}$) of the repeat length distribution at each locus. On average, ≈25 molecules were typed at each locus, and ≈500–1,000 molecules from 21–30 MS loci were typed from each tumor to characterize its genotype. For each tumor, we also summarize the variability among modal lengths ($\Delta_{germline}$) at different loci by their sample variance ($S^2_{loci}$).

The human tumors were first screened with up to 37 MS loci (Research Genetics, Huntsville, AL; list available on request). The 21–30 MS loci used for analysis were chosen for their amplification quality and a germline CA-repeat length of 16 or greater. The criterion of a minimum length repeat was based on published (21, 22) and unpublished observations that shorter repeat lengths tend to have lower mutation rates, and to avoid a potential lower length boundary constraint (see below).

**Modeling.** At the heart of our statistical approach is an algorithm for simulating the ancestry of a sample of cells (those used to estimate the genotype at each MS locus) taken from the tumor. We note that in our approach it is not necessary to simulate the history of the whole cell population that has been sampled. Once the ancestry of the sample has been simulated, mutations can be superimposed according to any mutation model [for example, a stepwise model (14, 23) in which loci accumulate random stepwise single repeat unit additions or deletions (reviewed in ref. 24)]. Thus, our method recreates both cellular division and mutation and the experimental process of sampling the final clonal expansion to ascertain the genotype ($\Delta_{germline}$) of each MS locus. The simulations start with the loss of MMR in a single cell with a germline genotype. A constant rate of 0.005 mutations per division [which is within the range of mutation rates observed in MMR-deficient cell lines (12, 13)] is assumed. Clonal expansion is defined as the time when this cell produces two daughter cells whose lineages persist. A tumor of approximately one billion cells is present at the end of each simulation.

To assess the sensitivity of the analysis to departures from the simple stepwise mutation model, we also used a model with range constraints (24) in which MS loci were not allowed to mutate to less than 12 repeats. For the range of estimated variances observed in Table 1 and starting (as in our experiments) with loci with lengths ≥16 repeats, the estimates of tumor age differ by at most 10% between the two scenarios (data not shown). Note that, if the mutation rate is halved, the estimated ages would be doubled. Simulation studies showed that age estimates obtained by sampling between 10 and 50 alleles or 20 and 30 loci varied little (data not shown).

**Estimating Tumor Histories.** Tumor histories were estimated by a computational inference method that involves matching the experimentally determined variances with the variances expected under a given model of cell division and MS mutation. A tumor history is defined by the time between loss of MMR and tumor removal and the time since clonal expansion (Fig. 1*a*). The simulations and theory indicate that, for a specific pattern of clonal expansion, the average value of $S^2_{alleles}$ is proportional to the time since initiation of the expansion. Therefore, the average of the experimental $S^2_{alleles}$ values obtained from all MS loci can be used to estimate the time since clonal expansion. Although expansion histories are likely to be variable for each tumor, simulations with a number of plausible scenarios (constant growth, immediate expansion followed by no growth, stepwise growth, and bottlenecks that reduce tumor populations up to 90%) indicate differences in average estimated ages of <20%. Therefore, we chose a model of constant growth to estimate the ages of the clonal expansions from $S^2_{alleles}$. For example, a cell is

replaced by an average of 1.061 cells in each division during the last 350 divisions of the adenoma of Patient I.

To estimate the age of a tumor, defined as the number of divisions, $g$, between sampling and loss of MMR, we use a method of moments approach. First, the simulation algorithm is used repeatedly to estimate how the expected value $E(S^2_{loci})$ of the variance $S^2_{loci}$ varies as a function of $g$; say, $E(S^2_{loci}) = f(g)$. For a given tumor with an observed variance of $s^2$, we estimate the age, $g$, by solving the equation $s^2 = f(g)$ to get the estimated age $g_{est}$. Once more, $f(g)$ is often an approximately linear function of $g$ over the range of interest, so finding $g_{est}$ is simple. Confidence intervals for $g_{est}$ may be found by a parametric bootstrap approach (cf. ref. 25, Section 2.2). We simulate a number, $B$, of replicates of the cell division, mutation, and sampling process, each running for $g_{est}$ divisions. For each of them, we calculate the variance among the loci and use this to reestimate the age of the process. This results in $B$ reestimates of the age: $g_1{}^*, \ldots g_B{}^*$, say. Without loss of generality, we can assume they are listed in increasing order. The values of $g_1{}^* - g_{est}, \ldots g_B{}^* - g_{est}$ are used to approximate the distribution of $g_{est} - g$. In particular, if $\alpha_1 < \alpha_2$ are such that $l = B\alpha_1$ and $m = B\alpha_2$ are integers, then a $100(1 - \alpha_1 - \alpha_2)\%$ confidence interval is given by $(2g_{est} - g_1{}^*, 2g_{est} - g_m{}^*)$.

## Results

In multistep progression (1–3), the final tumor arises from a single founder cell (Fig. 1*a*). The numbers of divisions before and after this founder cell are potentially reconstructed by a quantitative analysis using MS loci as molecular tumor clocks. Initiation is usually defined by an onset of neoplasia (1–3). However in this analysis, the start of progression is defined as the somatic loss of MMR because this event triggers an ≈100-fold (12, 13) increase in MS mutations (estimated here at 0.005 mutations per division). Therefore, the majority of mutations reflect divisions after loss of MMR, which begin to accumulate after only a few hundred divisions in MSI$^+$ tumors.

The quantitative analysis compares a tumor genotype with a starting genotype, which is likely identical to the germline genotype because normal tissues in our patients, as with most patients with MSI$^+$ tumors, lack detectable mutations (3, 26–28). Theory and simulations (see *Materials and Methods*) illustrate that how tumors progress influences their final patterns of MS mutations (Fig. 1). Simplistically, MS loci become polymorphic after clonal expansion because each cell accumulates mutations independently. The distribution of the tumor alleles is characterized by its variance ($S^2_{alleles}$). The difference in length between a tumor allele and its germline allele ($\Delta_{germline}$) is largely a function of the time before terminal clonal expansion. This is because multiple alleles in a clonal expansion mutate randomly and tend to drift around the founder allele rather than to drift coherently. Drift from germline (summarized at all loci by $S^2_{loci}$) predominately accumulates in the single cell lineage preceding the final clonal expansion.

The MS mutation patterns simulated under different progression scenarios are illustrated in Fig. 1*b*. If loss of MMR occurs at the time of terminal clonal expansion, drift from germline is small whereas tumor loci are polymorphic ($S^2_{loci}$ low, $S^2_{alleles}$ high). If loss of MMR occurs early relative to expansion, drift from germline is greater. This is further illustrated in Fig. 1*c* when identical expansion histories but different numbers of divisions preceding the final founder are simulated. $S^2_{loci}$ is a function of the number of divisions preceding terminal expansion. Therefore, we should be able to reconstruct the histories of individual MSI$^+$ tumors by sampling their MS allelic distributions and estimating $S^2_{loci}$ (proportional to the time since loss of MMR) and $S^2_{alleles}$ (proportional to the time since terminal clonal expansion). A rough simplification of our model is that an $S^2$ of ≈1 is equivalent to ≈200 divisions.

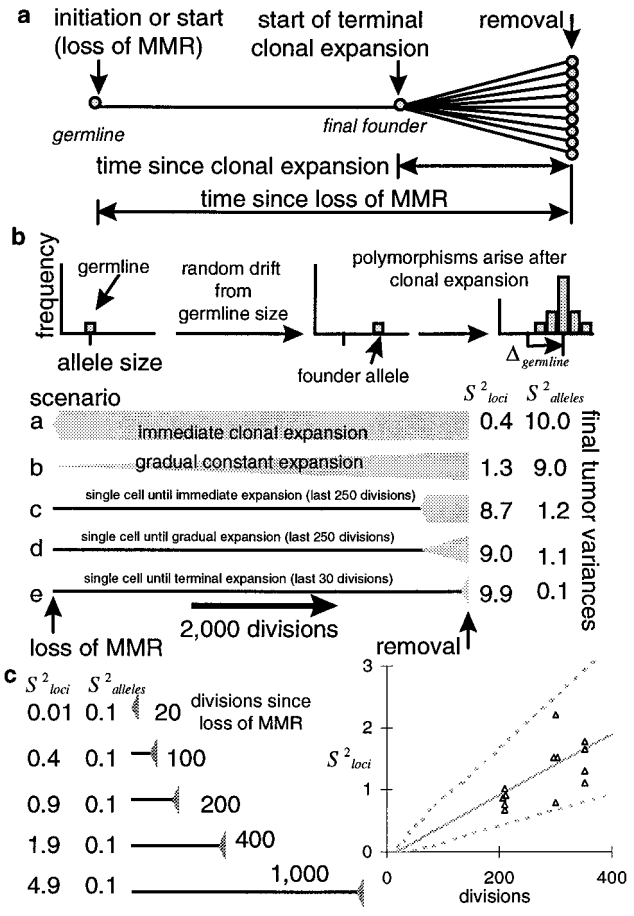


**Fig. 1.** (*a*) A tumor history is ''read'' from the mutations present in its cells. All of these cells are related to a single founder cell that represents the final bottleneck along a progression pathway. The history of this founder cell can be traced back along a single lineage because of the bottleneck nature of progression. We define the start of this pathway as the somatic loss of MMR. MS loci can record this history. Consider a single MS locus. After loss of MMR, the locus will randomly become larger or smaller, with the difference from its germline length a function of the number of divisions since loss of MMR. Upon terminal clonal expansion, the locus will become polymorphic. The time since terminal clonal expansion is reflected in the width or variance ($S^2_{alleles}$) of the tumor allele frequency distribution. The MS allele in the founder cell can be inferred by the most common allele present in the tumor. The time since loss of MMR preceding the bottleneck is reflected in the difference ($\Delta_{germline}$) between the length of the germline allele and the founder allele. Although the stochastic nature of mutation makes a single MS locus relatively uninformative, the analysis repeated at 20–30 different loci is robust. (*b*) Simulations of MS mutation. Data represents the results of 1,000 trials with 20–30 MS loci, and a symmetric stepwise model with the chance of addition of one repeat of 0.0025, and loss of one repeat of 0.0025, with the total mutation rate of 0.005 per division. In each scenario, the final tumor size is 1.0 cm$^3$ or one billion cells. Different patterns of MS mutations, summarized by the variances, $S^2_{alleles}$ and $S^2_{loci}$, are obtained with identical numbers of divisions (2,000) but different tumor histories. Therefore, a history of a human tumor can be inferred by sampling its MS alleles and estimating $S^2_{alleles}$ and $S^2_{loci}$. (*c*) Example of simulations with identical clonal expansion histories (20 terminal exponential divisions) but different times since loss of MMR. The average value of $S^2_{loci}$ increases with the numbers of divisions since loss of MMR whereas $S^2_{alleles}$ is constant. The graph illustrates the mean and 95% confidence intervals (dotted lines) of the simulations. The triangles in the graph represent values obtained from 210- to 352-day-old tissue culture experiments that mirror these simulations (see Fig. 2).

**Experimental Validation.** To test this model, we experimentally reconstructed progression pathways in which a single cell loses MMR and then subsequently undergoes clonal expansion at

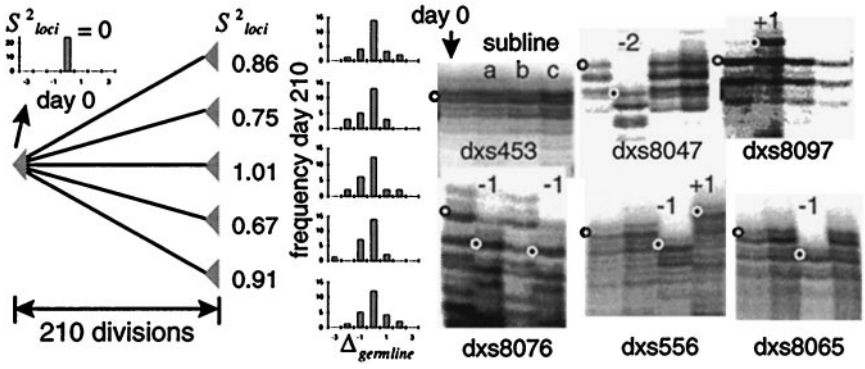


**Fig. 2.** A single HCT 116 cell was isolated, grown, and subsequently split into different sublines. After 190–332 days, single clones were isolated from each subclone, were expanded for 20 more divisions, and were typed at 24 MS loci (autoradiographs from three 210-day-old sublines and six loci are illustrated). Relative to germline (open circles or the allele size of the original clone), subline alleles are randomly the same, or larger or smaller (filled circles). The distribution of these changes ($\Delta_{germline}$) and their estimated variances ($S^2_{loci}$) are small compared with human tumors (compare with Fig. 3). Variation between alleles at a single locus is minimal because only 20 divisions occurred during terminal expansions ($S^2_{alleles} \approx 0$). This experiment mimics five simulated trials with a history of 210 divisions since loss of MMR and a terminal clonal expansion of 20 exponential divisions (see Fig. 1*c*). The experimental $S^2_{loci}$ values are consistent with the simulations (see triangles in the graph of Fig. 1*c*).

different times. A single clone of the MSI$^+$ colorectal cancer cell line HCT 116 was isolated and grown for 190–332 days as different sublines (Fig. 2). Single clones from the sublines were again isolated and then clonally expanded for 20 more divisions. Therefore, each of the expanded subclones have similar histories of late terminal clonal expansion after various days since "initiation." The MS genotypes of the final subline expansions were compared with the estimated genotype of the cell originally isolated at day zero. Consistent with the quantitative model, many of the loci had drifted from the "germline" lengths of the original clone (Fig. 2). The drift appeared random as each subline had alleles larger, smaller, or the same size as germline. For each subline, $S^2_{alleles}$ was essentially zero (data not shown) because clonal expansion was allowed for only 20 days. These control studies are consistent with our model as the subline $S^2_{loci}$ values, ranging from 0.6 to 2.2, were within estimated 95% confidence intervals (Fig. 1*c*).

This experimental model also illustrates the bottleneck nature of progression. The sublines were passaged as conventional cultures rather than maintained as single cells. However, the analysis only depends on starting from a single MMR deficient cell and subsequently selecting a single founder cell for terminal clonal expansion. Only the numbers of divisions before and the expansion history of the founder cell and not the expansion history before the founder cell affect the MS mutations present in the final expansion.

**Analysis of Human Tumors.** The MS alleles in the human MSI$^+$ colorectal tumors were sampled by diluting tumor DNA before PCR (Fig. 3). Approximately 10–30 alleles from 21–30 MS loci were analyzed for 13 tumors from nine patients (Table 1). Compared with the cell line studies, the tumor MS distribution modes had drifted more (Fig. 3), with estimated $S^2_{loci}$ values from 5.7 to 22.3 (Table 1). Estimated $S^2_{alleles}$ values were between 0.42 and 3.0. As a measure of consistency, $\Delta_{germline}$ and $S^2_{loci}$ from different parts of the same tumor should be similar and within simulated 95% confidence intervals because they presumably share identity by descent. This expectation was met (Fig. 3; data not shown).

The tumor histories derived from their MS mutations are illustrated in Fig. 4. The histories were different between tumors with an average of 2,300 divisions since loss of MMR repair and 280 divisions since clonal expansion. Assuming one division per day [a value consistent with intestinal stem cell studies (29)], the average age of the tumors was 6.2 yr since loss of MMR. Adenoma clonal expansion ages were slightly greater than cancer expansion ages (290 versus 280 divisions), and adenoma ages since loss of MMR were slightly less than the cancer ages (2,000 versus 2,400 divisions). However, histologic stages did not cor-

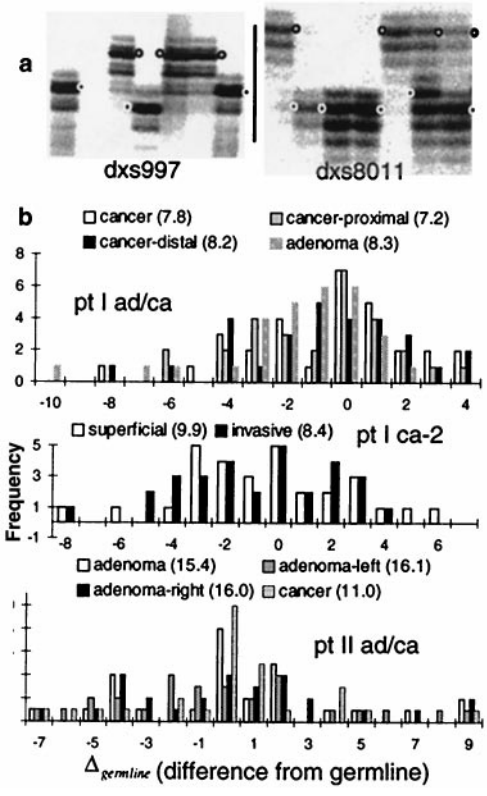


**Fig. 3.** (*a*) Autoradiographs of human tumor MS alleles. After dilution and PCR, the germline alleles (open circles) and tumor specific alleles (filled circles) become evident. Although the tumor alleles are polymorphic, they exhibit a modal size [−2 for DXS997 in the invasive cancer of Patient V (left) and −5 for DXS8011 in the right adenoma region of Patient II (right)]. (*b*) The differences from germline ($\Delta_{germline}$) of the 21–30 different MS loci from Patients I or II. Broader distributions, summarized by their variances ($S^2_{loci}$ in parentheses), indicate greater numbers of divisions since the loss of MMR (see Fig. 1 and compare with Fig. 2). Different regions from the same tumor have similar distributions and $S^2_{loci}$ values.

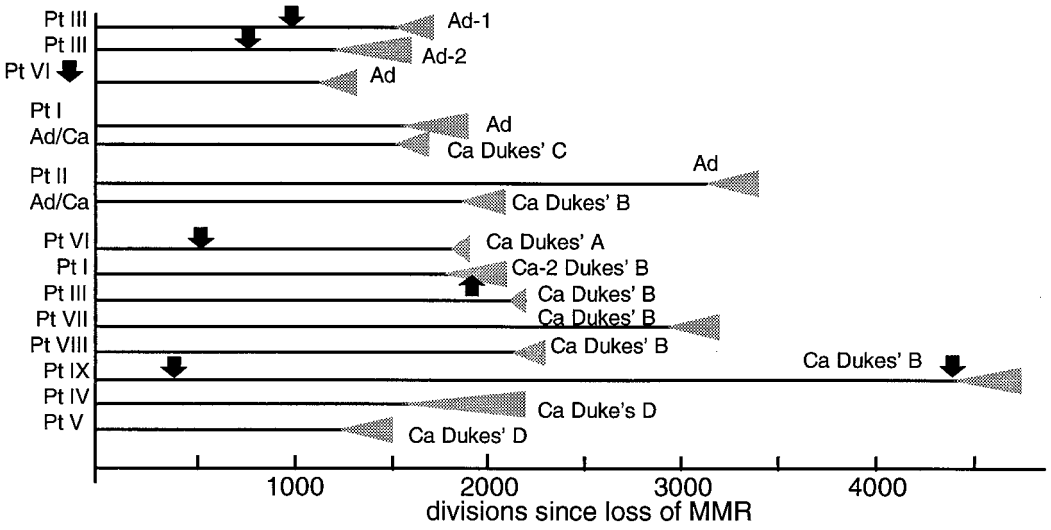


**Fig. 4.** Colorectal tumor histories inferred from their MS mutations. Each tumor history is unique, and most divisions occur before terminal clonal expansion (gray triangles). The average age is 2,300 divisions since loss of MMR. A trend toward increasing age with histologic progression is not apparent as the cancers and adenomas have similar ages. Negative clinical examinations (arrows, assuming one division per day) preceded six tumors, and for five tumors provided independent verification because they occurred before the postulated initiation of terminal clonal expansion. The negative examination in Patient I after the estimated initiation of the expansion of his Cancer-2 may indicate that it either was missed or was still too small to be detected. The data are consistent with genetic progression in the absence of visible progression, suggesting progenitor clone sizes are often below the threshold of clinical detection after loss of MMR.

respond to tumor ages as there was considerable variation between tumor types (Fig. 4).

Despite the variations in the tumor histories, the estimated ratios of divisions after terminal clonal expansion were relatively small (average 0.12) compared with the total divisions since loss of MMR. This finding suggests most divisions after loss of MMR occur before terminal clonal expansion. Note that our quantitative approach is limited to the division history of the final tumor and cannot recreate an entire progression history (such as an adenoma-cancer sequence) because it provides no information on clone sizes or phenotypes before the final founder cell. For example, the cancers may have arisen from and subsequently destroyed adenoma precursors or may have developed from occult precursors. Fortunately negative clinical examinations preceded six of our tumors, allowing independent verification of our model and knowledge of when a physically detectable precursor was probably absent.

The negative clinical examinations were consistent with the absence of clonal expansion inferred from five of our histories (Fig. 4). The history of the sole exception (Cancer-2 in Patient I) indicated terminal clonal initiated ≈170 days before the negative examination. However, this expansion may have been still too small to allow clinical detection. Under our scenario of constant growth, only ≈20,000 tumor cells would have been present at the time of the negative examination. In four of the tumors, the negative examinations occurred after loss of MMR but before terminal clonal expansion. Therefore, these tumors arising under clinical surveillance are more consistent with a scenario of occult progenitors rather than progression through a series of physically detectable clonal expansions.

## Discussion

Quantitative sequence comparisons have provided a powerful method for reconstructing population or species histories. Here we translate this approach to the problem of multistep tumor progression and use it to estimate numbers of divisions since loss of MMR and since a last clonal expansion. Estimated times had large confidence intervals, and other models of MS mutation and progression could yield alternative tumor histories. Nevertheless, our model follows the basic principles of multistep tumor progression (1), provides histories consistent with clinical and experimental observations, and accounts for the complex patterns of MS mutations observed in $MSI^+$ tumors with a simple mechanism of constant stepwise mutation.

Each of our tumors had a unique history, consistent with the hypothesis that tumor progression may follow different pathways (1, 2). The estimated intervals (Fig. 4) between initiation and removal (3.6–12.9 yr) were short, which may reflect the accelerated progression expected by a mutator phenotype (30). Despite this variability, one general feature was the early loss of MMR relative to terminal clonal expansion. The early loss of MMR is consistent with MS mutations distributed throughout $MSI^+$ tumors (12) and frameshifts in short repetitive sequences frequently found in their tumor suppressor loci (31–33).

Our tumor histories suggest that most progression occurs before terminal clonal expansion. Although this finding is not surprising for cancers, adenomas also had long progression intervals before clonal expansion and were of similar ages (averages of 2,000 versus 2,400 divisions) as the cancers. This suggests that adenomas arise after long periods of occult progression rather than their classical roles at the start of colorectal cancer progression. Adenomas may still contain precursors to $MSI^+$ cancers since the transition from adenoma to cancer may be rapid in the setting of MMR deficiency (7).

The negative surveillance examinations preceding some of our tumors also suggest progression occurs in the absence of visible manifestations. The emergence of a tumor shortly after a negative clinical examination may reflect a failure to detect a preexisting lesion, or rapid progression within the surveillance interval. Rapid progression seems unlikely because the interval tumors had similar numbers of MS mutations and estimated ages (6.0 versus 6.2 yr) as the tumors that arose in patients not under surveillance. Our analysis is more consistent with the absence of detectable tumors but the presence of occult progenitors at the times of the negative examinations. For example, although the physically detectable manifestations of the cancers of Patients I

and IX (Fig. 4) arose within less than 1 yr, the cancer progenitors had already accumulated >90% of their MS mutations at the times of their last negative clinical examinations.

It is important to note that histories derived from our quantitative analysis may differ from a physical record of progression. A visible record of progression requires selection and detectable clonal expansion. Therefore, some phases of progression may be invisible if their consequences do not immediately result in detectable physical changes. In contrast, we measure progression through MS mutations that can accumulate in the absence of selection, changes in phenotype, or detectable clonal expansion.

A critical question is whether the genetic progression measured by the accumulation of MS mutations is relevant to tumor progression. The time of the last clonal expansion can be directly related to a physically important progression milestone, but our studies suggest the early loss of MMR is unlikely associated with a recognizable progenitor. MS mutations can accumulate in phenotypically normal cells (8–11) so at least some of the progression recorded by MS loci may occur in overtly normal appearing crypts. Therefore, many of the divisions recorded by MS tumor clocks may occur before visible cycles of selection and clonal dominance. However, if cancer is a genetic disease, the accumulation of pertinent somatic mutations regardless of immediate selection and clonal expansion is relevant to progression, especially because neoplasia may not occur until a human cell has acquired a number of alterations (34). The loss of MMR appears to influence subsequent progression because many tumor suppressor loci such as APC, TGF-$\beta$ RII, Bax, and others have characteristic frameshift mutations in $MSI^+$ tumors (31–33). Although we measure mutations in noncoding MS loci, these mutations likely reflect a similar accumulation of mutations in loci that ultimately confer visible selection.

The current quantitative studies are consistent with multistep progression (1, 2). After loss of MMR, mutations gradually accumulate over time until the combinations sufficient for visible clonal expansion are acquired. The major difference with the classical adenoma–carcinoma sequence (3) is that much of progression occurs in clone sizes or phenotypes below a threshold of clinical detection. Although clonal expansion effectively increases mutation rates (35), expansion to a visible precursor may not be required as a prerequisite for $MSI^+$ progression because individual cells readily acquire mutations. At least in $MSI^+$ colorectal tumors, it appears that this previously occult prologue before visible neoplasia is longer than generally appreciated. The relatively long lengths of these prologues suggest the final phenotypes are contingent on many of the mutations acquired during occult progression. Given the technical and ethical problems with direct observations, quantitative analysis may provide the only feasible and objective method to reconstruct human tumor progression and probe even beyond the earliest physically detectable precursors. Additional studies with more or different types of loci should help further characterize the unique histories of individual tumors.

We thank C. Fajado, A. Arakawa, N. Yum, K. Adzhyan, C. B. Blake, and A. Ghatan for their technical assistance and Drs. Jeremy R. Jass and Barbara Leggett for providing appropriate specimens. This work was supported by grants from the National Institutes of Health to D.S. (CA58704 and CA70858) and National Science Foundation to S.T. (BIR95-04393).

1. Nowell, PC. (1976) *Science* **194,** 23–28.
2. Foulds, L. (1954) *Cancer Res.* **14,** 327–339.
3. Kinzler, K. W. & Vogelstein, B. (1996) *Cell* **87,** 159–170.
4. Otchy, D. P., Ransohoff, D. F., Wolff, B. G., Waver, A., Ilstrup, D., Carlson, H. & Rademacher, D. (1996) *Am. J. Gastroenterol.* **91,** 448–454.
5. Morson, B. C. (1974) *Cancer* **34,** 845–849.
6. Winawer, S. J., Zauber, A. G., Ho, M. N., O'Brien, M. J., Gottlieb, L. S., Sternberg, S. S., Waye, J. D., Schapiro, M., Bond, J. H., Panish, J. F., *et al.* (1993) *N. Engl. J. Med.* **329,** 1977–1981.
7. Jass, J. R. & Stewart, S. M. (1992) *Gut* **33,** 783–786.
8. Yao, X., Buermeyer, A. B., Narayanan, L., Tran, D., Baker, S. M., Prolla, T. A., Glazer, P. M., Liskay, R. M. & Arnheim, N. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 6850–6855.
9. Parsons, R., Li, G. M., Longley, M., Modrich, P., Liu, B., Berk, T., Hamilton, S. R., Kinzler, K. W. & Vogelstein, B. (1995) *Science* **268,** 738–740.
10. Ricciardone, M. D., Ozcelik, T., Cevher, B., Ozdag, H., Tuncer, M., Gurgey, A., Uzunalimoglu, O., Cetinkaya, H., Tanyeli, A., Erken, E. & Ozturk, M. (1999) *Cancer Res.* **59,** 290–293.
11. Wang, Q., Lasset, C., Desseigne, F., Frappaz, D., Bergeron, C., Navarro, C., Ruano, E. & Puisieux, A. (1999) *Cancer Res.* **59,** 294–297.
12. Shibata, D., Peinado, M. A., Ionov, Y., Malkhosyan, S. & Perucho, M. (1994) *Nat. Genet.* **6,** 273–281.
13. Bhattacharyya, N. P., Skandalis, A., Ganesh, A., Groden, J. & Meuth, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6319–6323.
14. Shibata, D., Navidi, W., Salovaara, R., Li, Z. H. & Aaltonen, L. A. (1996) *Nat. Med.* **2,** 676–681.
15. Shibata, D. (1997) *Am. J. Pathol.* **151,** 643–646.
16. Tsao, J. L., Tavaré, S., Salovaara, R., Jass, J. R., Aaltonen, L. A. & Shibata, D. (1999) *Am. J. Pathol.* **154,** 815–824.
17. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31,** 77–84.
18. Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. (1993) *Nature (London)* **365,** 274–277.
19. Umar, A., Boyer, J. C., Thomas, D. C., Nguyen, D. C., Risinger, J. I., Boyd, J., Ionov, Y., Perucho, M. & Kunkel, T. A. (1994) *J. Biol. Chem.* **269,** 14367–14370.
20. Lengauer, C., Kinzler, K. W. & Vogelstein, B. (1997) *Nature (London)* **386,** 623–627.
21. Weber, J. L. (1990) *Genomics* **7,** 524–530.
22. Hudson, T. J., Engelstein, M., Lee, M. K., Ho, E. C., Rubenfield, M. J., Adams, C. P., Housman, D. E. & Dracopoli, N. C. (1992) *Genomics* **13,** 622–629.
23. Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993) *Genetics* **133,** 737–749.
24. Goldstein, D. B. & Pollock, D. D. (1997) *J. Hered.* **88,** 335–342.
25. Davison, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Application* (Cambridge Univ. Press, Cambridge, U.K.).
26. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. (1993) *Nature (London)* **363,** 558–561.
27. Aaltonen, L. A., Peltomäki, P., Leach, F. S., Sistonen, P., Pylkkanen, L., Mecklin, J. P., Järvinen, H., Powell, S. M., Jen, J., Hamilton, S. R., *et al.* (1993) *Science* **260,** 812–816.
28. Thibodeau, S. N., Bren, G. & Schaid, D. (1993) *Science* **260,** 816–819.
29. Potten, C. S. & Loeffler, M. (1990) *Development (Cambridge, U.K.)* **110,** 1001–1020.
30. Loeb, L. A. (1991) *Cancer Res.* **51,** 3075–3079.
31. Huang, J., Papadopoulos, N., McKinley, A. J., Farrington, S. M., Curtis, L. J., Wyllie, A. H., Zheng, S., Willson, J. K. V., Markowitz, S. D., Morin, P., *et al.* (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9049–9054
32. Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L. Z., Lutterbaugh, J., Fan, R. S., Zborowska, E., Kinzler, K. W., Vogelstein, B., *et al.* (1995) *Science* **268,** 1336–1338.
33. Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J. C. & Perucho, M. (1997) *Science* **275,** 967–969.
34. Hahn, W. C., Counter, C. M., Lundberg, A. S., Beijersbergen, R. L., Brooks, M. W. & Weinberg, R. A. (1999) *Nature (London)* **400,** 464–468.
35. Tomlinson, I. P. M., Novelli, M. R. & Bodmer, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14800–14803.