American Journal of Pathology, Vol. 154, No. 6, June 1999

Colorectal Adenoma and Cancer Divergence

Evidence of Multilineage Progression

Jen-Lan Tsao,* Simon Tavaré,† Reijo Salovaara,‡ Jeremy R. Jass,§ Lauri A. Aaltonen,‖ and Darryl Shibata*

From the Department of Pathology, Norris Cancer Center, University of Southern California School of Medicine, Los Angeles, California; the Departments of Biological Sciences and Mathematics,† University of Southern California, Los Angeles, California; the Departments of Pathology and Medical Genetics,‡ Haartman Institute, University of Helsinki, Helsinki, Finland; the Department of Pathology,§ University of Queensland Medical School, Herston, Australia, and the Department of Medical Genetics,‖ Haartman Institute, University of Helsinki, Helsinki, Finland*

Colorectal cancer progression involves changes in phenotype and genotype. Although usually illustrated as a linear process, more complex underlying pathways have not been excluded. The object of this paper is to apply modern quantitative principles of molecular evolution to multistep tumor progression. To reconstruct progression lineages, the genotypes of two adjacent adenoma-cancer pairs were determined by serial dilution and polymerase chain reaction at 28–30 microsatellite (MS) loci and then traced back to their most recent common ancestor. The tumors were mismatch repair deficient, and therefore relatively large numbers of MS mutations should accumulate during progression. As expected, the MS genotypes were similar (correlation coefficients >0.9) between different parts of the same adenoma or cancer, but very different (correlation coefficients <0.2) between unrelated metachronous adenoma-cancer pairs. Unexpectedly, the genotypes of the adjacent adenoma-cancer pairs were also very different (correlation coefficients of 0.30 and 0.36), consistent with early adenoma-cancer divergence rather than direct linear progression. More than 60% of the divisions occurred after this early adenoma-cancer divergence. Therefore, the tumor phylogenies were not consistent with sequential stepwise selection along a single most "fit" and frequent lineage from adenoma to cancer. Instead, one effective early progression strategy creates and maintains multiple evolving candidate lineages, which are subsequently selected for terminal clonal expansion. ***(Am J Pathol 1999, 154:1815–1824)***

Multistep progression provides a unifying theme for carcinogenesis,[1–3] but its description may vary, depending on the perspective. A classical description is derived from the examination of mutation frequencies in tumors of different histological stages. For example, the adenoma-carcinoma sequence correlates larger and more dysplastic adenomas with the accumulation of greater numbers of mutations.[4,5] The object of this paper is to describe progression with an alternative and complementary approach. In this paper we apply principles of molecular evolution to infer past progression.

The study of evolution provides a useful analogy to illustrate some of the complex differences between descriptions based on the direct examination of presumptive precursors, or molecular evolution based on genetic comparisons between current species. The fossil record depicts physical aspects of evolution and provides a classical understanding of phylogenies, which have been complemented by modern molecular approaches. Phylogenies based on the fossil record or on genetic comparisons may not agree,[6] as they measure different aspects of evolution. One potential bias of the fossil record is its dependence on abundance, as rarer species may not be found. In contrast, molecular evolution can trace the divergence between species regardless of past abundance. In addition, the fossil record is replete with apparent dead ends, whereas genetic comparisons trace persistent lineages.

Similarly, multistep tumor progression may have multiple descriptions, depending on the approach and information desired. Progression models are largely based on mutation frequencies in lesions of different histological stages. By necessity this approach, similar to the analysis of fossils, is biased toward detectable clonal expansions, because one cannot physically analyze lesions that cannot be seen. The potential severity of this bias depends on (1) whether everything seen is relevant to cancer and (2) whether we see everything relevant to cancer. These issues are difficult to resolve. For example, clearly not

Supported by grants from the National Institutes of Health to D.S. (CA58704 and CA70858) and from the National Science Foundation to S.T. (BIR95-04393).

Accepted for publication March 10, 1999.

Address reprint requests to Dr. Darryl Shibata, Department of Pathology, Norris Cancer Center, University of Southern California School of Medicine, 1200 N. State St., Box 736, Los Angeles, CA 90033. E-mail: dshibata@hsc.usc.edu.

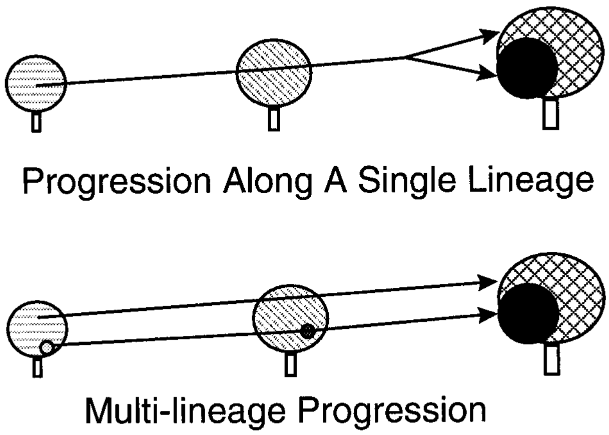


Figure 1. Two different pathways with the same visible manifestations. The description of progression is straightforward when it occurs along a single increasingly more "fit" and frequent lineage. Adenoma-cancer divergence occurs around the time of visible transformation. However, the situation is different if multiple lineages progress, especially if they do not expand before transformation. In this case the lineage destined for transformation diverged very early from the adenoma but remained physically occult until terminal transformation and detectable clonal emergence. It is worth noting that adenomas frequently harbor multiple clonal populations.[23,24] Although monolineage and multilineage progressions are fundamentally different, they may project similar macroscopic appearances. Bulk analysis of an adenoma may be misleading, as it will not reveal the presence or the significance of a rare precursor destined for transformation. However, genetic comparisons of adjacent adenoma-cancer pairs can distinguish between early or late lineage divergence.

every adenoma progresses, as the ratio of adenomas to cancers is approximately 30:1.[7] However, it is uncertain whether every adenoma could progress to cancer if allowed sufficient time.

Genetic comparisons of concurrent tumors can trace their ancestors regardless of past abundance. Mutation frequency studies and molecular tumor phylogenies should concur if progression occurs along a single, increasingly more "fit" and frequent lineage (Figure 1). In this case of sequential clonal evolution,[1] the clonal expansions are relevant to cancer (ie along the lineage to cancer), and the cells relevant to cancer expand (and therefore can be detected) before transformation. Physical (detectable) emergence coincides with lineage divergence.

However, genetic comparisons may conflict with a linear model if a lineage does not undergo detectable expansion before transformation, or if detectable expansions represent dead ends rather than direct precursors to cancer (Figure 1). In this case, multiple lineages related to a single precursor diverge early and independently undergo multistep progression either to adenoma or to cancer. Lineage divergence may precede detectable physical emergence by a long time if clonal expansion is contingent on mutations acquired after divergence. Although the implications of monolineage and multilineage progression are fundamentally different, mutation frequency studies are not explicitly designed to trace lineages and by default yield progression along a single lineage. Indeed, data from known multilineage progression (tumors from unrelated individuals) can be accommodated by a linear adenoma-cancer progression model.[4]

This paper traces tumor cell lineages back to their most recent common ancestor (MRCA) through comparisons between genotypes of concurrent tumor populations (Figure 2A). The position of a MRCA provides information on when distinct precursors are first created. As illustrated in Figure 2A, several MRCAs can be defined for the cells we examined. Comparisons between adjacent adenoma-cancer genotypes estimate the time since lineage divergence (T_2). Comparisons between different parts within the same tumor estimate the time since physical clonal emergence. Comparisons between germline and tumor genotypes estimate the time since initiation ($T_1 + T_2$).

Tumors originate from single progenitors that proliferate and generate many sublines or lineages that further branch or become extinct.[1] Lineages are dynamic and are defined by changes in fate rather than phenotype; a single lineage may historically project different phenotypes as it accumulates mutations. Although all cells within a tumor are related, tumor subpopulations may or may not be related by the same immediate lineage. Relationships between two subpopulations are illustrated with the trees in Figure 2B. With an early branch, subpopulations with different phenotypes can be defined as distinct lineages, because the MRCA has neither the "red" nor the "black" phenotype. In contrast, with a late branch lineages can be defined as related, because one subpopulation is derived from a cell with the phenotype of the other subpopulation, which is the phenotype of the MRCA. The new subpopulation is contingent on the first subpopulation—"black" cannot arise without "red" (Figure 2b).

The trees in Figure 2B represent a single transition. Tumor progression, however, is a multistep process[1–5] and therefore should be constructed as a series of sequential changes. Tumors change visible phenotypes, but may or may not change lineages with progression. For example, progression to cancer may occur along a single lineage (Figure 2C). Presumably, a new mutation in a single cell confers upon it a selective advantage, allowing clonal expansion and dominance over subclones that lack the mutation. Once past a "gatekeeper" mutation,[5] numerical predominance and greater numbers of accumulated oncogenic mutations could channel sequential selection along a single most "fit" and increasingly frequent lineage,[8,9] successively destroying and then shifting the MRCA to the right. If a cancer arises directly from a concurrent adenoma, adenoma-cancer lineage divergence should occur relatively late, as a late branch of the final step (Figure 1).

Alternatively, multiple related but subsequently independent lineages may persist, therefore preserving the MRCA (Figure 2D). Progression remains a multistep process, because each lineage independently undergoes sequential mutation, to either the final adenoma or final cancer phenotype. However, lineage divergence can occur early and may precede the visible differentiation of cancers out of adenomas. In this case, the adenoma and cancer lineages are distinct because they branch early and the MRCA had neither the adenoma nor cancer phenotype. The adenoma lineage could be "erased"

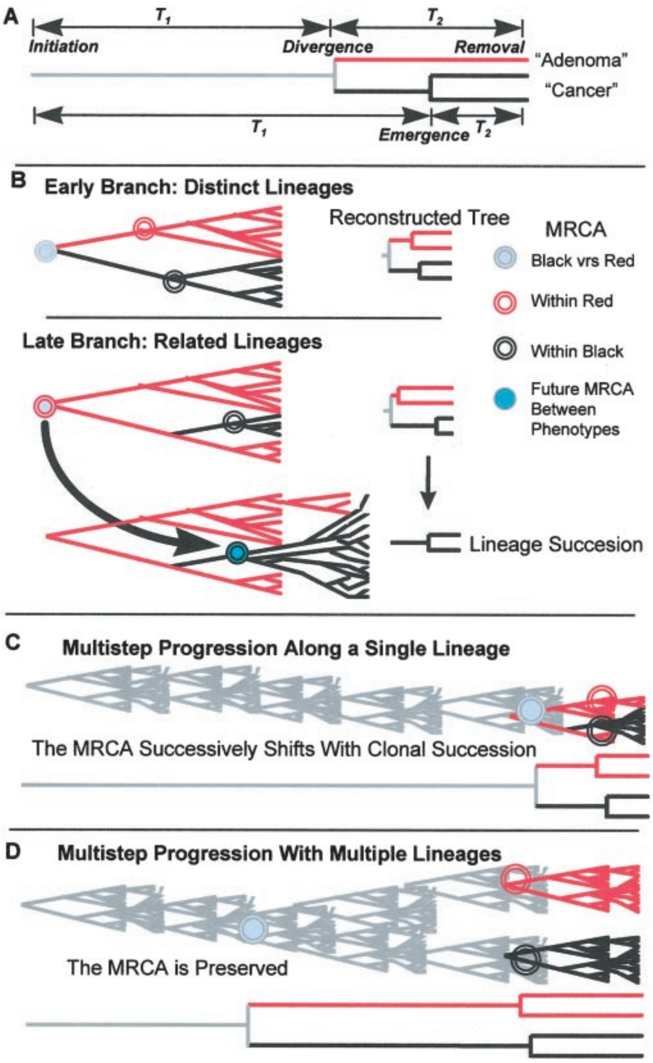

A
T_1
T_2
Initiation
Divergence
Removal
"Adenoma"
"Cancer"
T_1
Emergence
T_2
B
Early Branch: Distinct Lineages
Reconstructed Tree
MRCA
Black vrs Red
Within Red
Within Black
Future MRCA Between Phenotypes
Late Branch: Related Lineages
Lineage Succesion
C
Multistep Progression Along a Single Lineage
The MRCA Successively Shifts With Clonal Succession
D
Multistep Progression With Multiple Lineages
The MRCA is Preserved

Figure 2. A: Reconstructed tumor trees. Different events can be estimated by comparing genotypes from different tumor regions. The period of identity by descent with a common lineage is T_1. The time since divergence (T_2) is estimated from differences between adenoma and cancer regions and indicates their MRCA or the point at which their distinct precursors were created. The time since physical emergence is estimated through comparisons between different parts within the same tumor. The time since initiation ($T_1 + T_2$) is estimated from differences between tumor and germline genotypes. Although lineages can be traced, their phenotypes may change through time and are only known at removal. B: The relationship between two subpopulations can be defined by the phenotype of their MRCA. Lineages can be defined as distinct if they branch early and their MRCA does not share the phenotype of either lineage. In this case, neither lineage is a direct precursor of the other. In contrast, lineages are related if they branch later, as the MRCA has the phenotype of one of the lineages. In this case, one subpopulation arises directly from a single cell with the phenotype of the other subpopulation. After lineage succession with either late or early branching, a new MRCA will be created to the right of the former MRCA when another distinct phenotype arises. C: Multistep tumor progression represents a series of transitions, or combinations of the early and late branches from B. With progression along a single lineage, a mutation in a single cell confers a selective advantage, leading to clonal expansion and dominance. Subsequent mutations in expanded subclones lead to serial stepwise clonal evolution along an increasingly more "fit" and frequent lineage. Phenotype and genotype but not lineage change with progression. Each step successively shifts the MRCA to the right. D: Multilineage, multistep progression. Lineages persist, and therefore the MRCA remains fixed. When these lineages differentiate or physically emerge, their genotypes may be very different, because they diverged early. Therefore, the transition from the adjacent adenoma to the cancer involves changes in phenotype, genotype, and lineage. Many variations are possible, and only some of the dead ends are illustrated. Regardless of complexity, adjacent adenoma-cancer genotypes can be compared and then traced to a MRCA. The phenotypes and their clone sizes (which may be as few as one cell) during progression are unknown; they are illustrated in gray in C and D.

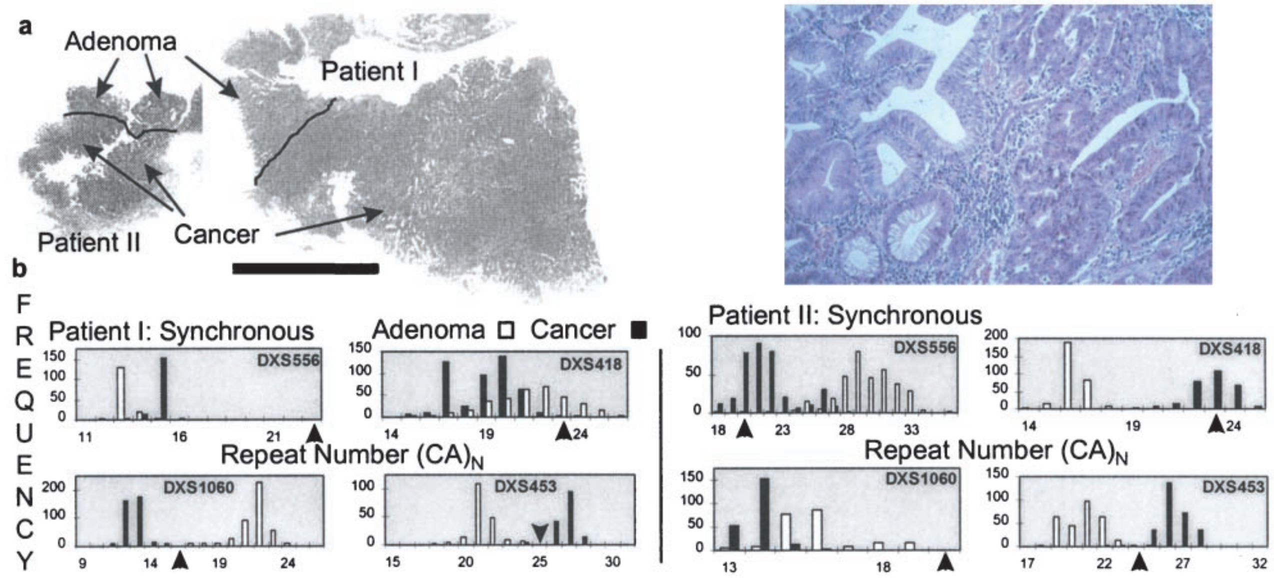


Figure 3. a: Outlines of the synchronous and adjacent adenoma-cancer pairs. The photomicrograph illustrates the junction of the adenoma (left) and cancer of patient II (H&E, ×100). b: Frequency distributions of MS alleles from the adenoma and cancer regions. The distributions illustrate relatively limited heterogeneity within an adenoma or a cancer. At some loci, the distributions are distinct between the adjacent adenoma-cancer pairs. Arrowheads indicate germline alleles.

Table 1. Characteristics of the Synchronous and Metachronous Tumors

Patient	Age*	Type	Tumor	Stage	Interval†	Estimated age‡
I	43	HNPCC	Adenoma/Cancer-1	Dukes' C	—	1600
			Cancer-2	Dukes' B	6 months	1800
II	42	Sporadic	Adenoma/cancer	Dukes' B	—	2600
III	43	HNPCC	Cancer	Dukes' B	—	2000
			Adenoma	0.5 cm	10 years	1500

*Age at first tumor.
†Interval between the first and second metachronous tumors.
‡Estimated number of divisions since initiation (loss of MMR) and removal. Ages are averages from Table 3.

back to the MRCA without changing the cancer. It is worth noting that phenotypes and clone sizes (which may be as few as a single cell) during early progression are unknown; these are illustrated in gray in Figure 2C and D.

Genotypes can be compared and traced to a MRCA. Tumors deficient in DNA mismatch repair (MMR) have greatly elevated mutation rates,[5] which are most prominent in simple repeat sequences or microsatellites (MS). With mutation rates as high as 0.01 per division,[10,11] noncoding MS loci in mutator phenotype (MSI+) tumors can function as "molecular tumor clocks" because they are expected to become polymorphic after relatively few divisions.[12,13] MS loci mutate predominantly by a relatively predictable mechanism ("slippage" during DNA replication[14,15] with small repeat unit additions or deletions,[10,11] allowing linkage of lineages through time. MS loci have been used to trace the emergence of modern humans out of Africa and the divergence between humans and chimpanzees.[16–18]

Physically adjacent colorectal adenomas and cancers are presumptive evidence that cancers arise from adenomas.[19] Comparisons of their MS genotypes allow objective analysis of this relationship. When do adjacent adenomas and cancers diverge?

Materials and Methods

Specimens

Five tumors from three patients were examined (Table 1). Patients I and III were from hereditary nonpolyposis colorectal cancer (HNPCC) families and had germline mutations in hMS2 and hMLH1, respectively. Although patient II was only 42 years old at the time of his colorectal cancer, his familial history did not meet the criteria for HNPCC. There were two synchronous and adjacent adenoma-cancer pairs, and two physically distinct metachronous adenoma-cancer pairs.

Sampling

Individual cells cannot be readily isolated and genotyped from fixed tissues. Therefore, the essential approach isolates the DNA from specific tumor regions on a thin 4–8-μm-thick microscope slide. The isolated DNA comes from a mixture of adjacent cells of similar phenotype and is fragmented in such a way that MS loci are physically separated from each other. DNA is sampled at random from this pool, and essentially single loci are typed by polymerase chain reaction (PCR) after dilution. The process is repeated from the same pool until alleles from multiple loci are typed.

MS Typing

To simplify analysis, X-chromosomal CA-dinucleotide repeat MS loci and male patients were used. Every allele therefore represents a single cell because MSI+ tumors characteristically lack aneuploidy.[20] The MS distributions were determined by two methods. For DXS556, DXS1060, DXS418, and DXS453 (Research Genetics, Huntsville, AL) and the data of Figure 3, multiple small tumor regions of approximately 200–400 cells and containing at least 70% tumor cells were isolated by selective ultraviolet radiation fractionation[12] from microscopic tissue sections. The DNA in these dots was diluted to essentially single alleles[12] with about 20–80% of reactions yielding PCR products, which were analyzed on 6% denaturing polyacrylamide sequencing gels and a phosphoimager (Molecular Dynamics, Sunnyvale, CA). PCR products were labeled with [^{33}P]dCTP (NEN Research Products, Boston, MA) incorporated during 38–43 PCR cycles in 5-μl reaction volumes.

Normal cell contamination was minimized by subtracting germline alleles by truncation when tumor alleles were different from germline. When tumor MS distributions appeared to include germline-sized alleles, 30% of all alleles were considered to originate from contaminating normal cells and were subtracted from the germline allele frequency. The MS allelic distributions were similar between tumor dots within a cancer or adenoma and therefore were combined to yield the composite distributions of Figure 3.

For the data of Figure 4 and Table 2, DNA was extracted in bulk from dissected adenoma or cancer regions and then diluted for analysis. At least 10 alleles were amplified for each additional MS locus (list available on request) until a mode became evident.

Results

The strategy compares MS genotypes (CA-repeat unit number) between synchronous and physically adjacent MSI+ adenoma-cancer pairs (Table 1 and Figure 3a). The MS loci were polymorphic, and their distributions were different be-

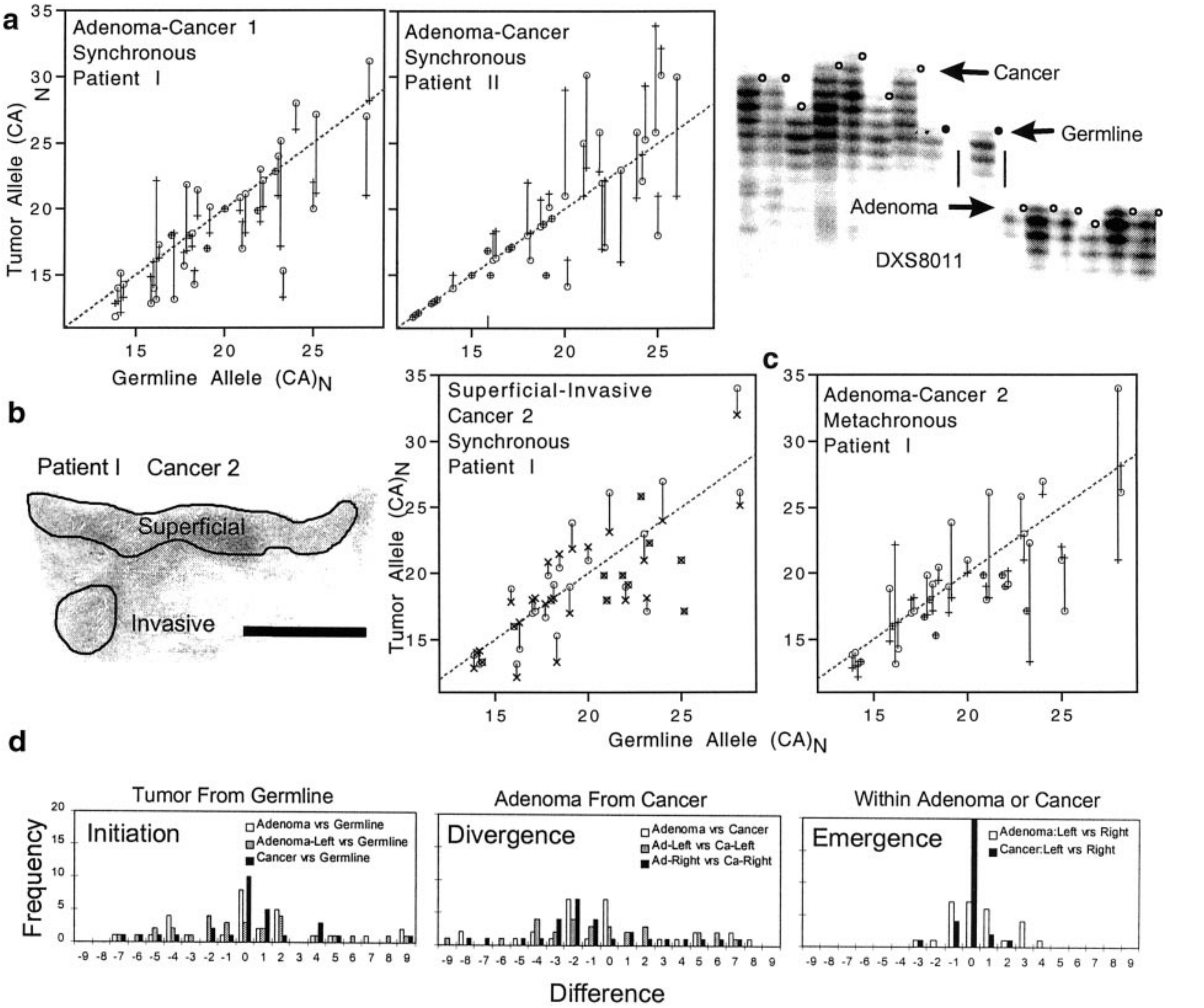


Figure 4. a: MS tumor modes from 34–37 different loci. The solid vertical lines connect the adenoma (cross-hatches) and cancer (open circles) modes at each locus and represent relative genetic distances from each other and from the germline. A representative autoradiograph from patient II illustrates the difference between the adenoma and adjacent cancer (final estimated modal sizes, respectively, −4 and +4 repeat units from the germline $(CA)_{26}$ allele). The genotypes were different between the adjacent adenoma-cancer pairs, suggesting distant rather than direct relationships. b: Outline of the superficial and invasive portions of Cancer-2 of patient I, which arose within a surveillance interval of 6 months. The modes of the superficial (open circles) and invasive portions (crosses) of Cancer-2 were more alike compared to the adjacent adenoma-cancer pairs. c: MS modes of the unrelated metachronous adenoma-cancer pair from patient I. Similar to the adjacent adenoma-cancer pairs, the genotypes of the metachronous tumors were different. d: Alternative display of some of the data from patient II. Differences between tumor and germline genotypes (time since initiation) are greater than within tumors (time since physical clonal emergence). Differences between adenoma and cancer genotypes are also large, suggesting early lineage divergence.

tween the adjacent tumors (Figure 3b). The adenoma-cancer pairs were further compared at a total of 34–37 loci. These additional loci were not characterized as extensively as in Figure 3b. Specifically, the ability to amplify single alleles consistently at every locus was uncertain, and fewer alleles were genotyped (~25 alleles per locus). The modes of the loci were evident from their distributions (see, for example, Figure 3b) and would be less susceptible to problems of reduced PCR sensitivity. Therefore, for experimental simplicity, the MS distributions for all of the loci are summarized by their modes, which are assumed to approximate their means (see Appendix). The genotypes defined by these modes differed between the adjacent adenoma-cancer pairs (Figure 4a).

Although physical proximity has been used to infer that adenomas are direct precursors to cancers,[19] the MS analysis suggests a more distant relationship. To further understand how quickly adenomas and cancers diverge, several other scenarios were examined. Different regions within the same tumor should be closely related, because they presumably arise from the same terminal clonal expansion. For example, the invasive and superficial portions of a cancer arising within a surveillance interval of 6 months had very similar genotypes (Figure 4b). In contrast, metachronous tumors arising at different sites and times in the same patient should be unrelated, because they initiate from the same germline but otherwise progress independently. Metachronous adenoma-cancer pairs exhibited large MS differences similar to adjacent adenoma-cancer pairs (Figure 4c).

The data were modeled assuming a single cell initiates tumorigenesis through loss of MMR and later splits into two different tumor lineages (see Appendix). Using the

Table 2. Relative Times Since Divergence

Patient	Type	Tumor pair	No. of loci*	ρ	Confidence intervals	Relative time since divergence[†]
I	S (Synchronous)	Adenoma: Cancer-1	30	0.30	(−0.06, 0.60)	0.70
	S	Left: right (Cancer-1)	30	0.96	(0.92, 0.98)	0.04
	S	Superficial: invasive (Cancer-2)	30	0.91	(0.81, 0.96)	0.09
	M (Metachronous)	Adenoma: Cancer-2	30	0.11	(−0.26, 0.45)	0.89)
	M	Cancer-1: Cancer-2	30	0.14	(−0.23, 0.48)	0.85
II	S	Adenoma: Cancer	28	0.36	(−0.02, 0.65)	0.64
	S	Left: right (Adenoma)	28	0.91	(0.81, 0.96)	0.09
	S	Left: right (Cancer)	28	0.96	(0.91, 0.98)	0.04
III	M	Adenoma: Cancer[‡]	28	0.16	(−0.24, 0.51)	0.84
	S	Left: right Adenoma	25	0.96	(0.91, 0.98)	0.04

Relative times from divergence are calculated from the estimated correlation coefficient ρ between MS lengths at a given locus in the two tissue types, using formula (5).

*Loci with germline alleles less than 16 repeats were not included for analysis, because Figure 4 suggests such short alleles are relatively more stable.

[†] $T_2/(T_1 + T_2)$, where $T_1 + T_2$ is the time since loss of MMR and T_2 is the time since lineage divergence.

[‡] One locus had a very high expansion size relative to germline and appears to be an outlier. If this point is included in the analysis, we obtain the estimate $\rho = 0.49$ with a 95% confidence interval of (0.14, 0.73). If this single point is removed, the correlation is estimated to be 0.16, with a 95% confidence interval of (−0.24, 0.51). Neither estimate changes the overall conclusion that synchronous adenoma-cancer pairs are not closely related.

model, we estimate the relative numbers of divisions between initiation, divergence, and presentation (Table 2 and Figure 5). The estimated intervals between initiation and presentation of the synchronous tumor pairs were between 1600 and 2600 divisions, or 4.4–7.1 years, assuming one division per day (Table 1). As expected, unrelated metachronous adenoma-cancer pairs from patients I and III diverged very early. In contrast, different regions within the same tumor were closely related and diverged later. Less than 10% of their divisions occurred after their MRCA, illustrating a random although predictable accumulation of MS mutations in both adenoma and cancer cells.

Adjacent adenoma-cancer lineages diverged relatively early (Figure 5). The adjacent pairs appeared to be related because they diverged later than the metachronous pairs, but still evolved independently for greater than 60% of their divisions. However, as recently observed for some adenomas,[21] the adjacent tumors may also be unrelated because the estimated confidence intervals were large (Table 2). For patient II, adenoma-cancer lineage divergence substantially preceded adenoma or cancer intratumor divergence (Figures 4d and 5).

Discussion

Clonal evolution[1] encompasses a large number of possible pathways to cancer. Genetic comparisons of adjacent adenoma-cancer pairs potentially reconstruct a number of critical steps. Correlation between genotypes at a single locus can be used to estimate the time since initiation (tumor versus germline genotypes), the time since lineage divergence (adenoma versus cancer genotypes), and the time since physical clonal emergence (different parts within the same tumor).[12] For each comparison, the fewer the intervening divisions, the greater the correlation between genotypes. For example, in Figure 4d the differences in modal genotype are plotted for each MS locus. The larger spread evident in the comparisons between adenoma and cancer, as opposed to the smaller spread observed with comparisons within adenoma or within cancer, is consistent with earlier divergence.

Although the data and their analysis are complex, they yield relatively simple tumor trees (Figure 5). Some aspects of these trees are known with certainty and therefore can be examined for internal consistency. As expected, the unrelated metachronous adenoma-cancer pairs had very different genotypes (correlation coefficients <0.2) and diverged early. Different regions within the same adenoma or cancer had more similar genotypes (correlation coefficients >0.9) and therefore had the recent divergence expected of a clonal expansion. The estimated divergence (160 days) between superficial and invasive portions of Cancer-2 is within the 6-month clinical surveillance interval of patient I.

Adjacent adenoma-cancer pairs also had very different genotypes (correlation coefficients of 0.30 and 0.36). Therefore their lineages diverged relatively early and evolved independently for more than 60% of their ~1600–2600 divisions since initiation. The visible manifestations at adenoma-cancer divergence are unknown because lineages and not phenotypes are traced. Phe-

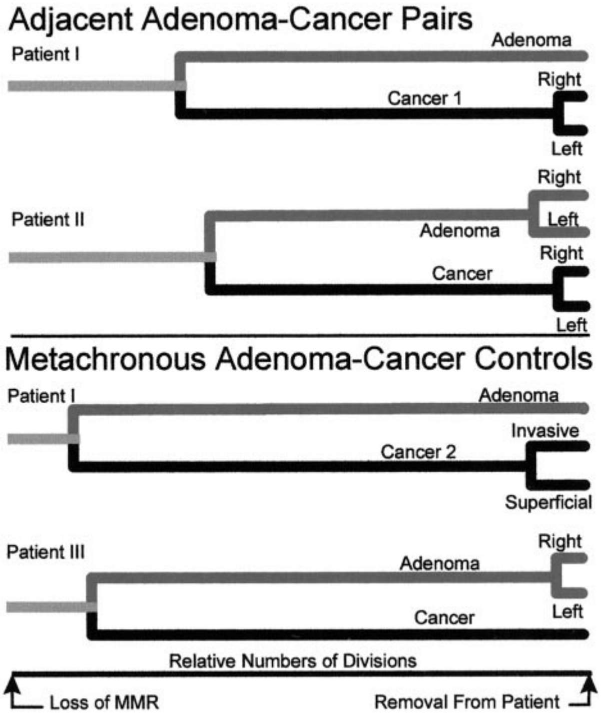


Figure 5. Adenoma-cancer lineages. The tumor trees are potentially more complex because only terminal expansions are analyzed and many other occult lineages may be present. The total branch lengths are normalized. With both metachronous and synchronous tumors, adenoma and cancer lineages progressed independently for the majority (>60%) of their evolution. The 95% confidence intervals for the estimated periods before the splits (T_1) between all of the adenoma-cancer pairs include zero, so it is possible that, similar to the metachronous pairs, the synchronous pairs were never directly related. In contrast, different regions within the same adenoma or cancer are more closely related and are divergent for less than 10% of their evolution. Multiple comparisons between and within adjacent tumors (patient II) illustrate that adenoma-cancer lineage divergence occurred substantially before physical adenoma or cancer emergence. Rather than selection of a single most "fit" lineage, multiple lineages are created, maintained, and then subsequently selected for terminal clonal expansion.

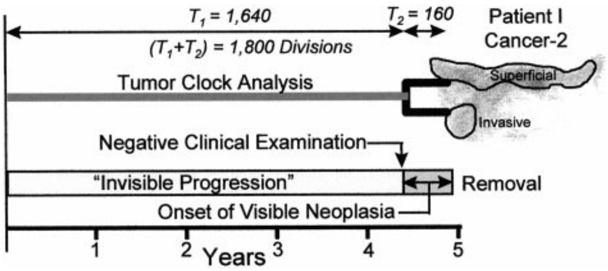


Figure 6. Comparison between the tumor clock analysis and clinical surveillance. Cancer-2 appeared to develop very quickly, as a physical precursor was not visible 6 months before it was removed. The clock analysis, however, suggests that an "invisible" precursor was present, which had already accumulated ~90% of the MS mutations present in the final cancer during the years since initiation (assuming one division per day). The superficial and invasive portions of the cancer appeared to physically emerge shortly after the negative clinical examination. This example illustrates that most mutations accumulate before transformation, but not necessarily in a visible precursor.

notypic differentiation or detectable emergence of the first distinct adenoma or cancer precursors may coincide with lineage divergence or, more likely, may be delayed and contingent on the subsequent accumulation of mutations. For example, the adjacent tumors of patient II, with an estimated age of 2600 divisions since loss of MMR, physically emerged recently, but their lineages diverged earlier (Figure 5). Since divergence occurred substantially before detectable emergence, it is unlikely that the two cells created at the lineage split from their MRCA had yet acquired their respective adenoma or cancer phenotypes.

Cancer-2 of patient I also provides insight into the period before detectable emergence, because no physical precursor was visible 6 months before it was removed. Although evolution in this cancer may have been compressed, with accelerated formation and subsequent destruction of an adenoma precursor, its MS genotype was very different from its germline. With an estimated age of 1800 divisions (Table 1) since loss of MMR, a more likely scenario is that most mutations (~90% if cells divide once a day) accumulated in an occult or microscopic precursor (Figure 6). Therefore, Cancer-2 and early adenoma-cancer lineage divergence suggest that a cancer lineage may not require macroscopic expansion before transformation. Adenoma precursors may not be essential in the setting of a mutator phenotype because a high mutation rate can compensate for the lack of clonal expansion.[3,22] Therefore, a lineage selected for visible adenoma expansion may not coincide with the pathway to cancer,[7] as other underlying occult lineages may have greater potential for transformation. Neoplasia becomes a consequence of rather than an obligate substrate or "direction" for further random mutation, because some expansions represent dead ends with respect to the final cancer lineage.

The current study does not directly address whether the cancer precursor had an adenoma phenotype, because it traces lineages and not their variable phenotypes. Although it is possible that the cancers may have destroyed a more closely related adenoma or more extensive sampling may uncover adenoma regions more related to the cancers, the concurrent adenoma and cancer lineages still diverged early. Therefore, transformation likely involved a switch to a related but distinct lineage rather than extension along the adjacent adenoma lineage. Although the possibility of a sudden burst in genomic instability with transformation[23] cannot be excluded, such an increase would appear to be temporary, as most cancers do not exhibit significant mutational heterogeneity,[23,24] whereas our mechanism is specific to MSI+ tumors and postulates a constant, high mutation rate. Occult multilineage progression with periodic physical emergence of different lineages, like progression along a single lineage, can account for observations that adenomas have fewer mutations than cancers and that adjacent adenoma-cancer pairs often exhibit mutational differences.[4,5,8,23,25]

Early divergence requires the prolonged persistence of multiple lineages, which implies physical protection or niches against any temporally dominant subclones. The crypt architecture and stem cell renewal of normal intestinal mucosa provide protection against neoplastic expansion.[26] Recent studies of MSI+ adenomas[13,27] demonstrate evidence of stem cell behavior, which would inherently limit clonal succession.

The current study reconstructs MSI+ tumor lineages to peer back in time to the first point at which a cancer branches or attains a destiny distinct from that of a concurrent adenoma. This critical event may be occult and apparent only in retrospect. Although refinement of the model is possible, the substantial genetic differences between adenoma-cancer pairs (in contrast to the intratumor similarities) would be difficult to generate unless their lineages were separated by large numbers of divisions. Multilineage progression may be more universal than currently appreciated, as lineages are seldom formally traced. Unlike the predictable visible chronology of linear multistep progression (sequential clonal expansion), this study lessens the potential bias imposed by the analysis of only visible clonal expansions and therefore emphasizes the length and complexity of steps that may not be directly observed. Without observable criteria (such as clonal frequency) to order progression, one cannot predict which lineages are destined for transformation, except in retrospect. Perhaps multilineage progression reflects a similar biological inability to select early on the most likely candidate for transformation. Instead of clonal evolution through serial stepwise selection of a single most "fit" and frequent lineage from adenoma to cancer, one effective early progression strategy creates and maintains multiple evolving candidate lineages, which are subsequently selected for terminal clonal expansion.

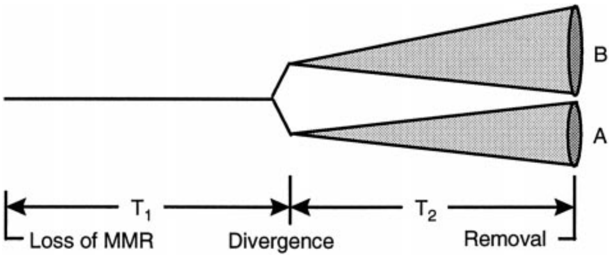
Acknowledgments

We thank A. Arakawa, N. Yum, K. Adzhyan, and C. B. Blake for their technical assistance and Drs. Heikki J. Järvinen, Jukka-Pekka Mecklin, and Barbara Leggett for providing appropriate specimens.

Appendix

Theory

For synchronous tumors, we assume a cell alive at generation $T_1 - 1$ splits into two cells in generation T_1, one of which is the precursor to the A cell population, the other the B cell population. The A and B cell populations evolve independently of each other for a further T_2 divisions: After these T_2 divisions a number of cells are sampled from cell lineage A and from lineage B. In our experimental approach, we are able to measure the modal repeat length at each locus. In what follows we use these modal repeat lengths as surrogates for the average length in each sample (see Figure 3). The average lengths $\bar{L}_A$ and $\bar{L}_B$ of a given MS locus in lineages A and B, respectively, are measured relative to germline. To calculate the correlation coefficient ρ between $\bar{L}_A$ and $\bar{L}_B$, we use a stepwise model of MS evolution.[28–32] Under this model, a random number of repeats are added at the given locus at each cell division. These repeat numbers are independent from division to division, with common mean μ and variance σ^2. (For simplicity we assume the same parameters hold in all divisions, although our approach can be modified to account for different mutation mechanisms at different stages at the expense of greater complexity.)



We denote by S_A the random number of divisions taken to trace two cells chosen at random from the A cell population back to their MRCA. In our model, we must have $1 \leq S_A \leq T_2$. The mean of S_A is denoted by $E(S_A)$. Under this mutation model, it is clear that

$$E(\bar{L}_A) = (T_1 + T_2)\mu. \tag{1}$$

Two cells chosen at random from population A share a number of ancestral cells (those in generations 0,1, . . ., T_1, $T_1 + 1$, . . ., $T_1 + T_2 - S_A$) and therefore share part of their evolutionary history. After time $T_1 + T_2 - S_A$, however, the two cells have independent mutation histories. We can use this observation to show that the variance of $\bar{L}_A$ is given by

$$\text{var}(\bar{L}_A) = \sigma^2[T_1 + (T_2 - E(S_A)) + E(S_A)/n_A], \tag{2}$$

where n_A is the number of cells sampled from lineage A. The results in Equations (1) and (2) apply to the B lineage, with S_A replaced by S_B (the random time taken to trace two B lineage cells back to their MRCA), and n_A by n_B (the number of cells sampled from the B lineage).

On the other hand, if we choose a cell at random from each of lineages A and B, they will share fewer ancestral cells, those in generations 0,1, . . ., $T_1 - 1$. After time T_1, the two cells have independent mutation histories. Using this fact, we can show that the covariance between $\bar{L}_A$ and $\bar{L}_B$ is given by

$$\text{cov}(\bar{L}_A,\bar{L}_B) = \sigma^2(T_1 - 1). \tag{3}$$

Combining Equations (2) and (3) and simplifying shows that

$$\rho = \text{corr}(\bar{L}_A,\bar{L}_B) = \frac{T_1 - 1}{(T_1 + T_2)c'} \tag{4}$$

where

$$c = \sqrt{1 - \frac{(n_A - 1)E(S_A)}{n_A(T_1 + T_2)}}\ \sqrt{1 - \frac{(n_B - 1)E(S_B)}{n_B(T_1 + T_2)}}.$$

We note from Equation (4) that the correlation coefficient is always positive, and that $\rho = 0$ in the particular case where $T_1 = 1$ (corresponding to independent evo-

Table 3. Tumor Means, Variances, and Ages (Appendix)

Patient	Tumor	No. of loci*	Mean	Variance	SE (Mean)	Estimated ages†
I	Adenoma	30	−1.47	8.33	0.53	1700
	Cancer-1	30	−0.57	7.77	0.51	1600
	Cancer-1, left	29	−0.69	7.22	0.50	1400
	Cancer-1, right	29	−0.79	8.17	0.53	1600
	Cancer-2 superficial	30	−0.47	9.91	0.58	2000
	Cancer-2 invasive	30	−0.87	8.46	0.53	1700
II	Adenoma	28	0.50	15.4	0.74	3000
	Cancer	28	0.19	11.0	0.63	2200
	Adenoma, left	28	−0.11	16.1	0.76	3200
	Adenoma, right	28	0.32	16.0	0.76	3200
	Cancer, left	28	−0.54	9.52	0.58	1900
	Cancer, right	28	−0.57	8.99	0.57	1800
III	Adenoma	29	0.14	10.1	0.59	2000
	Cancer	29	−1.00	10.3	0.60	2000
	Adenoma, left	25	−0.44	6.34	0.50	1300
	Adenoma, right	25	−0.32	6.23	0.50	1200

*Loci with germline alleles less than 16 repeats were not included for analysis because Figure 4 suggests that such short alleles are relatively more stable.

†This column estimates the number of divisions between initiation and clinical presentation and is based on the measured variances for each tumor and a mutation rate of 0.005 per division (see above). A lower mutation rate would correspondingly increase the estimated ages.

lution of two cell lineages for time T_2). Furthermore, ρ is expected to be small whenever T_1 is small relative to T_2.

To obtain a simpler, approximate formula for ρ, we assume that the total number of cell divisions $T_1 + T_2$ is large relative to $E(S_A)$ and $E(S_B)$. It follows that $c \approx 1$, so that

$$\rho \approx \frac{T_1}{T_1 + T_2.} \tag{5}$$

Estimation of Tumor Ages

The method employed here estimates the size of T_2 relative to $T_1 + T_2$. To estimate the absolute size of $T_1 + T_2$, we need to assume something more about the mutation mechanism. For example, if mutations arise according to the simplest symmetrical mutation model (in which with probability P a mutation occurs, and results in the addition or loss of a single repeat unit, each with equal chance 0.5), then $\mu = 0$ and $\sigma^2 = P$. With the possible exception of the adenoma in patient I, the data are consistent with the assumption $\mu = 0$. Knowledge of P then allows us to estimate $T_1 + T_2$, using the variances calculated in Equation (2). This follows because

$$\mathrm{var}(\bar{L}_A) \approx p[T_1 + T_2 - E(S_A)] \approx p(T_1 + T_2), \tag{6}$$

assuming as earlier that $E(S_A)$ is much smaller than $T_1 + T_2$. The result in Equation (6) is employed to give, in the final column of Table 3, estimates of the likely numbers of divisions necessary to generate the observed MS mutations for each tumor. As an example, from Table 3 the variance of patient II adenoma is estimated to be approximately 15. Hence if $P = 0.005$ (a value within the range of mutation rates observed in MMR deficient cell lines),[10,11] it follows that $T_1 + T_2 \approx 15/0.005 = 3000$ divisions.

Statistical Analysis

For synchronous tumor pairs, we can apply the theory directly to MS variability measured (relative to germline) at several different loci. We use metachronous tumor pairs as controls, treating them as though they started from the same cell. Because these pairs have evolved independently after initiation, we would expect that the correlation between the two mean lengths to be zero; this is reflected in the fact that the confidence interval for ρ should include zero.

Our model makes specific predictions for the variability that may be used informally to assess the adequacy of the model. In particular, Equations (1) and (2) show that $E(\bar{L}_A)$ and $E(\bar{L}_B)$ are equal, and that the variances $\mathrm{var}(\bar{L}_A)$ and $\mathrm{var}(\bar{L}_B)$ are equal. Estimates of these means and variances, obtained from the different loci and A,B pairs, are given in Table 3.

For the synchronous tumor pairs, formal statistical tests of equalities of means and variances are complicated because of the weak dependence between the different MS loci (they share part of the same cell lineage history). Assuming approximate independence of the loci, the results in Table 2 show no obvious contradictions with the model: the tumor pairs have approximately the same means and variances. This is also true of the metachronous pairs.

The correlation estimates given in Table 1 are also based on comparison of $\bar{L}_A$ and $\bar{L}_B$ over different loci, for each pair of tumors A,B. The confidence intervals 33 for

ρ are again based on the assumed adequacy of the approximate independence of the ($\bar{L}_A$, $\bar{L}_B$) pairs over different loci within a given individual.

References

1. Nowell PC: The clonal evolution of tumor cell populations. Science 1976, 194:23–29
2. Foulds L: The experimental study of tumor progression: a review. Cancer Res 1954, 14:327–339
3. Loeb LA: Mutator phenotype may be required for multistage carcinogenesis. Cancer Res 1991, 51:3075–3079
4. Fearon ER, Vogelstein B: A genetic model for colorectal tumorigenesis. Cell 1990, 61:759–767
5. Kinzler KW, Vogelstein B: Lessons from hereditary colorectal cancer. Cell 1996, 87:159–170
6. Gibbons A: Genes put mammals in age of dinosaurs. Science 1998, 280:675–676
7. Koretz RL: Malignant polyps: are they sheep in wolves' clothing? Ann Intern Med 1993, 118:63–68
8. Vogelstein B, Kinzler KW: The multistep nature of cancer. Trends Genet 1993, 9:138–141
9. Tomlinson IPM, Novelli MR, Bodmer WF: The mutation rate and cancer. Proc Natl Acad Sci USA 1996, 93:14800–14803
10. Bhattacharyya NP, Skandalis A, Ganesh A, Groden J, Meuth M: Mutator phenotypes human colorectal carcinoma cell lines. Proc Natl Acad Sci USA 1994, 91:6319–6323
11. Shibata D, Peinado MA, Ionov Y, Malkhosyan S, Perucho M: Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. Nature Genet 1994, 6:273–281
12. Shibata D, Navidi W, Salovaara R, Li ZH, Aaltonen LA: Somatic microsatellite mutations as molecular tumor clocks. Nature Med 1996, 2:676–681
13. Shibata D: Molecular tumor clocks and dynamic phenotype. Am J Pathol 1997, 151:643–646
14. Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M: Frameshift mutations and the genetic code. Cold Spring Harb Symp Quant Biol 1966, 31:77–84
15. Strand M, Prolla TA, Liskay RM, Petes TD: Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 1993, 365:274–277
16. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: High resolution of human evolutionary trees with polymorphic microsatellites. Nature 1994, 368:455–457
17. Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW: Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 1995, 92:6723–6727
18. Deka R, Jin L, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, Ferrell RE, Chakraborty R: Population genetics of dinucleotide (dC-dA)$_n$ • (dG-dT)$_n$ polymorphisms in world populations. Am J Hum Genet 1995, 56:461–474
19. Sugarbaker JP, Gunderson LL, Wittes RE: Colorectal cancer. Cancer: Principles and Practices in Oncology. Edited by VT DeVita Jr, S Hellman, and SA Rosenberg. Philadelphia, Lippincott, 1985, pp 800–803
20. Lengauer C, Kinzler KW, Vogelstein B: Genetic instability in colorectal cancers. Nature 1997, 386:623–627
21. Novelli MR, Williamson JA, Tomlinson IPM, Elia G, Hodgson SV, Talbot IC, Bodmer WF, Wright NA: Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. Science 1996, 272:1187–1190
22. Loeb LA: Microsatellite instability: marker of a mutator phenotype in cancer. Cancer Res 1994, 54:5059–5063
23. Borland CR, Sato J, Appelman HD, Bresalier RS, Feinberg AP: Microallelotyping defines the sequence and tempo of allelic losses at tumor suppressor gene loci during colorectal cancer progression. Nature Med 1995, 1:902–909
24. Shibata D, Schaeffer J, Li ZH, Capella G, Perucho M: Genetic heterogeneity of the c-K-*ras* locus in colorectal adenomas but not adenocarcinomas. J Natl Cancer Inst 1993, 85:1058–1063
25. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL: Genetic alterations during colorectal-tumor development. N Engl J Med 1988, 319:525–532
26. Cairns J: Mutation selection and the natural history of cancer. Nature 1975, 255:197–200
27. Tsao JL, Zhang J, Salovaara R, Li ZH, Järvinen HJ, Mecklin JP, Aaltonen LA, Shibata D: Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. Am J Pathol 1998, 153:1189–1200
28. Valdes AM, Slatkin M, Freimer NB: Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 1993, 133:737–749
29. Shiver MD, Jin L, Chakraborty R, Boerwinkle E: VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 1993, 134:983–993
30. Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. Genetics 1995, 139:457–462
31. Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW: An evaluation of genetic distances for use with microsatellite loci. Genetics 1995, 139:463–471
32. Kimmel M, Chakraborty R: Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor Popul Biol 1996, 50:345–367
33. Snedecor GW, Cochran WG: Statistical Methods, 7th ed. Ames, IA, Iowa State University Press, 1980, pp 186ff