[CANCER RESEARCH 63, 4577–4582, August 1, 2003]

The Impact of Microsatellite Instability on the Molecular Phenotype of Colorectal Tumors[1]

Yuriko Mori,[2] Florin M. Selaru,[2] Fumiaki Sato, Jing Yin, Lisa A. Simms, Yan Xu, Andreea Olaru, Elena Deacu, Suna Wang, Jennifer M. Taylor, Joanne Young, Barbara Leggett, Jeremy R. Jass, John M. Abraham, David Shibata, and Stephen J. Meltzer[3]

Department of Medicine, Division of Gastroenterology [Y. M., F. M. S., F. S., J. Y., Y. X., A. O., E. D., S. W., J. M. A., S. J. M.] and Department of Surgery [D. S.], University of Maryland School of Medicine and Greenebaum Cancer Center and Baltimore VA Hospital, Baltimore, Maryland 21201; Conjoint Gastroenterology Laboratory, Queensland Institute of Medical Research, Queensland 4029, Australia [L. A. S., J. Y., B. L.]; Queensland Centre for Schizophrenia Research, The Park Centre for Mental Health, Queensland 4076, Australia [J. M. T.]; and Department of Pathology, McGill University, Montreal, Quebec H3A 2B4, Canada [J. R. J.]

ABSTRACT

Frequent microsatellite instability MSI (MSI-H) occurring in human tumors is characterized by defective DNA mismatch repair and unique clinical features. However, infrequent MSI (MSI-L) has not been attributable to any other defined molecular pathway, and its existence as a biologically distinct category has been challenged. Moreover, the global molecular phenotypes (GMPs) underlying MSI-H, MSI-L, or microsatellite-stable (MSS) tumors have never been evaluated. To evaluate the impact of MSI status on GMP and to determine the importance of MSI relative to other molecular and clinical features, cDNA microarray-derived data from 41 colon cancers were interpreted using principal components analysis. The clinically relevant principal component with the greatest impact on GMP was component 3, which distinguished MSI-H from non-MSI-H (*i.e.*, MSI-L and microsatellite stable) tumors and was designated the MSI-H separator. Notably, MSI-L cancers were also clearly distinguished from non-MSI-L tumors by another principle component, component 10 (the "MSI-L separator"). This second finding validates the existence of MSI-L tumors as a distinct molecular phenotypic category. Thus, both components 3 and 10 reflected different aspects of MSI and helped to establish principal components analysis as a useful tool to identify and characterize distinct biological features of human malignancy.

INTRODUCTION

MSI-H,[4] due to defective DNA MMR, identifies a unique group of human cancers with distinct clinicopathological features. Ten to 15% of colorectal cancers exhibit MSI-H and are reported to have unique clinical characteristics, including poor differentiation, severe inflammatory cell infiltration, proximal anatomic location, a favorable response to chemotherapy, and a good prognosis relative to MSS tumors (1–6). Disruption of the DNA MMR system in sporadic MSI-H cancers is most often caused by somatic promoter methylation of the MMR gene *MLH1* (7–10), whereas germ-line mutation in the MMR genes *MLH1* or *MSH2* is the most common cause of MSI in the familial colon cancer syndrome hereditary nonpolyposis colon cancer (HNPCC; Refs. 11–13). The carcinogenic pathway underlying MSI-H cancers (the MSI pathway) is considered distinct from the chromosomal instability pathway underlying non-MSI-H cancers (14, 15). For example, in contrast to non-MSI-H cancers, aneuploidy and mutations of *RAS* and *TP53* are rare in MSI-H cancers (16–18). However, comprehensive molecular phenotypes underlying the MSI pathway have not yet been fully clarified.

Non-MSI-H cancers are currently classified into two subcategories: MSS cancers and cancers with low-level MSI (MSI-L cancers). However, MSI-L cancers lack the definitive molecular and biological features observed in MSI-H cancers; therefore, the current definition of MSI-L cancers is based purely on MSI frequency (19). Thus, opinions differ regarding the existence of the MSI-L category because both the interpretations of low MSI frequency and the cutoff value for this low frequency are poorly defined (20). Nevertheless, some researchers consider MSI-L tumors to be a biologically distinct category, with unique molecular features (16, 21). For example, loss of expression of the *MGMT* gene has been suggested as a potential underlying molecular mechanism for MSI-L tumors (22). However, most non-MSI-H cancers exhibit extremely low frequencies of MSI when a large number of microsatellite loci are tested (23, 24). It is not clear how many of these non-MSI-H cancers actually belong in the MSI-L category, and how many are actually MSS. In addition, several studies have failed to reveal significant differences in clinicopathological features between MSS and MSI-L cancers (17, 25).

Data derived from cDNA microarrays can be analyzed using two sets of bioinformatics approaches: unsupervised and supervised. When microarray data are analyzed using unsupervised techniques, these data can define new natural biological categories, independent of biases introduced by preexisting classifications (such as MSI-H, MSI-L, and MSS). To date, the most widely used unsupervised analytic method has been hierarchical clustering (26, 27). Although results from cluster analysis are easily interpreted, this method often fails to distinguish between subtle subcategories of lesions or between relevant and irrelevant gene expression data (28). PCA is a different unsupervised approach. In contrast to clustering, PCA can discover multiple layers of meaning within microarray data. PCA searches for key variables or components in a multidimensional data set to explain differences among observations (29). The components extracted are independent, allowing for mining of data in a layer-oriented fashion. Thus, after one particular component has been identified and analyzed for its possible significance, it is possible to isolate its influence and move to the next component or layer of information (30). In addition, PCA provides a quantitative measure of the fraction of sample variance generated by each component to the variance in the whole data set.

Our goals in the current study were to evaluate the influence of MSI status and other clinicopathological characteristics on global molecular phenotypic data and to test whether these data support the existence of MSI-L as a distinct category of tumors. Therefore, we performed cDNA microarray analyses of 41 primary colorectal cancers and applied the unsupervised technique, PCA, to reveal whether any natural subgroupings of these tumors corresponded to their MSI status.

Received 1/10/03; accepted 5/2/03.

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

[1] This research was supported by Public Health Service Awards CA85069, DK47717, CA77057, CA95323, CA01808, and CA98450 and by the Office of Medical Research, Department of Veterans Affairs.

[2] These authors contributed equally to this article.

[3] To whom requests for reprints should be addressed, at 655 West Baltimore Street, BRB 8-009, Baltimore, MD 21201. Phone: (410) 706-3375; Fax: (410) 706-1099; E-mail: smeltzer@medicine.umaryland.edu.

[4] The abbreviations used are: MSI-H, frequent microsatellite instability; MMR, mismatch repair; MSS, microsatellite stable; MGMT, *O*-6-methylguanine-DNA methyltransferase; MSI, microsatellite instability; MSI-L, low-level microsatellite instability; MSS, microsatellite stable; PCA, principal components analysis; aRNA, amplified RNA; GMP, global molecular phenotype; FDR, false discovery ratio.

Table 1 *Correlations between PCA components and tumor characteristics*

This table displays the p-values of each correlation between a component (leftmost column) and a tumor characteristic (MSI status, clinical characteristics, and experimental batch, top row). Results are shown for the thirteen components with the highest impact on global molecular phenotype, each accounting for more than 2% of total data variability (rightmost column). Student *t*-test and one-way ANOVA test were used as statistical tests for characteristics with two subgroups and with three or more subgroups, respectively. For each tumor characteristic, the method of grouping tumors and the number of cases in the relevant group is shown (second row, *Grouping for p-value calculation*; H, MSI-H; L, MSI-L; S, MSS; *MD*, moderately differentiated; *MP*, moderately to poorly differentiated; *PD*, poorly differentiated). One of the tumors lacked the information for differentiation, location, Dukes stage, and lymph node metastasis and was excluded from analyses regarding these characteristics. Additional six tumors lacked the lymph node metastasis information and were exluded from analyses of this characteristic. The percent variance (i.e., the fractional influence on gene expression profile) exerted by each component is shown in the *rightmost* column. Significant component-clinical trait correlations, i.e., those with p-values less than 0.01 and 0.05, are indicated as described in the footnote. Components 3 and 10 correlated significantly with MSI status, while component 6 correlated significantly with Dukes stage and lymph node metastasis.

Categories	MSI status			Differentiation		Location	Dukes stage	Lymph node metastasis	Gender	Age Correlation	*P*	Experimental batch	% variance
Grouping for *P* calculation	H (12) L (14) S (15)	H (12) L + S (29)	L (14) H + S (27)	MD (26) MP (7) PD (7)	MD (26) MP + PD (14)	L (19) R (21)	A (3) B (22) C (11) D (4)	+ (12) − (22)	M (21) F (20)			A (13) B (13) C (15)	
Component 1	0.6294			0.0252[a]		0.0073[b]	0.1576	0.0156[a]	0.6777	−0.079	0.6265	<0.0001[b]	16.16
Component 2	0.4568			0.7963		0.1678	0.5870	0.1489	0.2743	0.173	0.2813	<0.0001[b]	12.82
Component 3	<0.0001[b]	<0.0001[b]	0.0230[a]	0.0014[b]	0.0003[b]	0.0043[b]	0.2604	0.7306	0.8521	0.181	0.2582	0.3700	7.04
Component 4	0.3592			0.5451		0.7885	0.7179	0.4442	0.0459[a]	−0.023	0.8855	0.0886	6.24
Component 5	0.6452			0.2720		0.0110[a]	0.8807	0.7486	0.1256	0.192	0.2300	0.6062	4.53
Component 6	0.0999			0.8194		0.4847	0.0036[b]	0.0016[b]	0.6925	0.075	0.6421	0.6630	3.87
Component 7	0.6287			0.0346[a]		0.5441	0.4729	0.1545	0.3083	0.108	0.5023	0.8302	3.27
Component 8	0.8420			0.7058		0.7504	0.6875	0.1692	0.1685	0.182	0.2570	0.9580	2.87
Component 9	0.9553			0.3044		0.7052	0.8521	0.9109	0.3980	−0.162	0.3127	0.8896	2.67
Component 10	0.0122[a]	0.1025	0.0030[b]	0.1652		0.9703	0.6304	0.9000	0.5004	0.380	0.0137[a]	0.9862	2.49
Component 11	0.9630			0.6285		0.6607	0.3386	0.1154	0.2448	0.133	0.4086	0.9393	2.33
Component 12	0.7855			0.7111		0.5458	0.6603	0.3861	0.5385	−0.114	0.4788	0.9904	2.11
Component 13	0.5090			0.8655		0.9362	0.5577	0.8063	0.8605	−0.290	0.0648	0.9380	2.02

[a] $P < 0.05$.
[b] $P < 0.01$.

The current study was designed to assess the impact of MSI on GMP and to rank this impact relative to other clinicopathological features. This study suggests that MSI-H exerts a dominant influence on GMP. Furthermore, MSI-L *per se* has not been attributable to defective MMR, nor has it been defined on a molecular basis or firmly established as a distinct tumor category. The global molecular phenotypic data described below support the existence of MSI-L tumors as a distinct biologic category.

MATERIALS AND METHODS

Colorectal Tumors and MSI Status Testing

Forty-one primary colorectal cancers from our tissue archives were studied, including 12 MSI-H, 14 MSI-L, and 15 MSS colon cancer specimens. Primary colorectal tumor and their corresponding normal colorectal mucosa specimens were obtained at surgery at the University of Maryland Medical System and Royal Brisbane Hospital. Genomic DNAs and total RNAs were extracted from fresh frozen specimens following the methods described previously (31, 32). Classification of MSI status was based on 11 microsatellite markers (BAT25, BAT26, BAT40, D2S123, D5S346, D10 S197, D17S250, D18S34, D18S55, MYBT22, MYCL, ACTC, and BAT34C4) described in a National Cancer Institute Workshop in 1998 (19). BAT25, BAT26, and BAT40 were mononucleotide repeats, whereas D2S123, D5S346, D10S197, D17S250, D18S34, D18S55, MYBT22, ACTC, and BAT34C4 were dinucleotide repeats. MYCL was a tetranucleotide repeat. Tumors completely lacking MSI were labeled MSS, those with MSI in >30% of informative loci were designated MSI-H, and those with MSI in at least 1 but <30% of informative loci were designated MSI-L. All MSI-H tumors showed MSI in at least one mononucleotide marker, whereas all MSI-L tumors showed MSI only in dinucleotide or tetranucleotide repeat markers. Primer sequences for MSI status assessment are available on line.[5] A detailed PCR protocol has been described elsewhere (33).

cDNA microarrays

Preparation of the aRNA Probe. aRNA was amplified from 20–50 μg of total RNA with a T7-based protocol (34, 35). Labeling was performed on 3–6 μg of aRNA by incorporating Cy3- or Cy5-labeled dCTP using random primers and Superscript reverse transcriptase (28). The reference probe was prepared from an equimolar mixture containing aRNAs from the eight human malignant cell lines: HCT116; HT29; CaCo-2; HCT15 (colon cancer); HTB114 (leukemia); MCF-7 (breast cancer); HeLa (cervical cancer); and AGS (gastric cancer). These cell lines were included in the reference probe according to our previous microarray work (36–39).

Microarray Preparation and Hybridization. Microarray slides containing 8064 sequence-verified human cDNA clones were prepared according to a previously described protocol (28). The Lawrence Livermore Laboratories cDNA library was used as a clone source (Invitrogen, Carlsbad, CA). All clones were independently sequence verified in our own laboratory. Microarrays were cohybridized to a Cy5-labeled specimen aRNA and to Cy3-labeled reference probe (28) at 65°C overnight. After hybridization, each slide was scanned using a GenePix 4000A dual laser slide scanning system (Axon Instruments, Union City, CA).

Data Analysis

Preprocessing Gene Selection. We included in analysis only clones with expression information for at least 96% of the tumors (*i.e.*, clones lacking information in only two or fewer tumors). This minimal information threshold was surpassed by 6242 of 8064 printed clones.

Data Preprocessing. Datapoints representing gene expression ratios were log transformed. We then normalized data to exclude intensity-dependent bias. In this fashion, local distortions in signal and background intensity within different regions of a slide were overcome. We based this procedure on the assumption that Cy5/Cy3 ratios should not depend on spot intensity. This type of data distortion was removed by a robust scatterplot smoothing method (40). Using SigmaPlot version 5 (SPSS, San Rafael, CA), we calculated the Lowess fitting curve using a fitting parameter of 40%.

Hierarchical Agglomerative Clustering. Data imported from GenePix was manipulated and clustered, using established algorithms implemented in the software program Cluster (26, 41). Average linkage clustering with centered correlation was used. TreeView software (26, 41) generated visual representations of clusters.

PCA. All PCA-related calculations were performed in MatLab (MathWorks, Inc., Natick, MA). The data, filtered as described above, were input into MatLab and normalized so that for each specimen, the mean gene expression was 0 and the SD was 1. The number of independent dimensions in PCA is equal to the number of specimens minus one. Thus, because there were 41 specimens in our study, 40 independent components were derived. The

[5] Internet address: http://microarray.umaryland.edu/manuscripts/MoriCAN/.

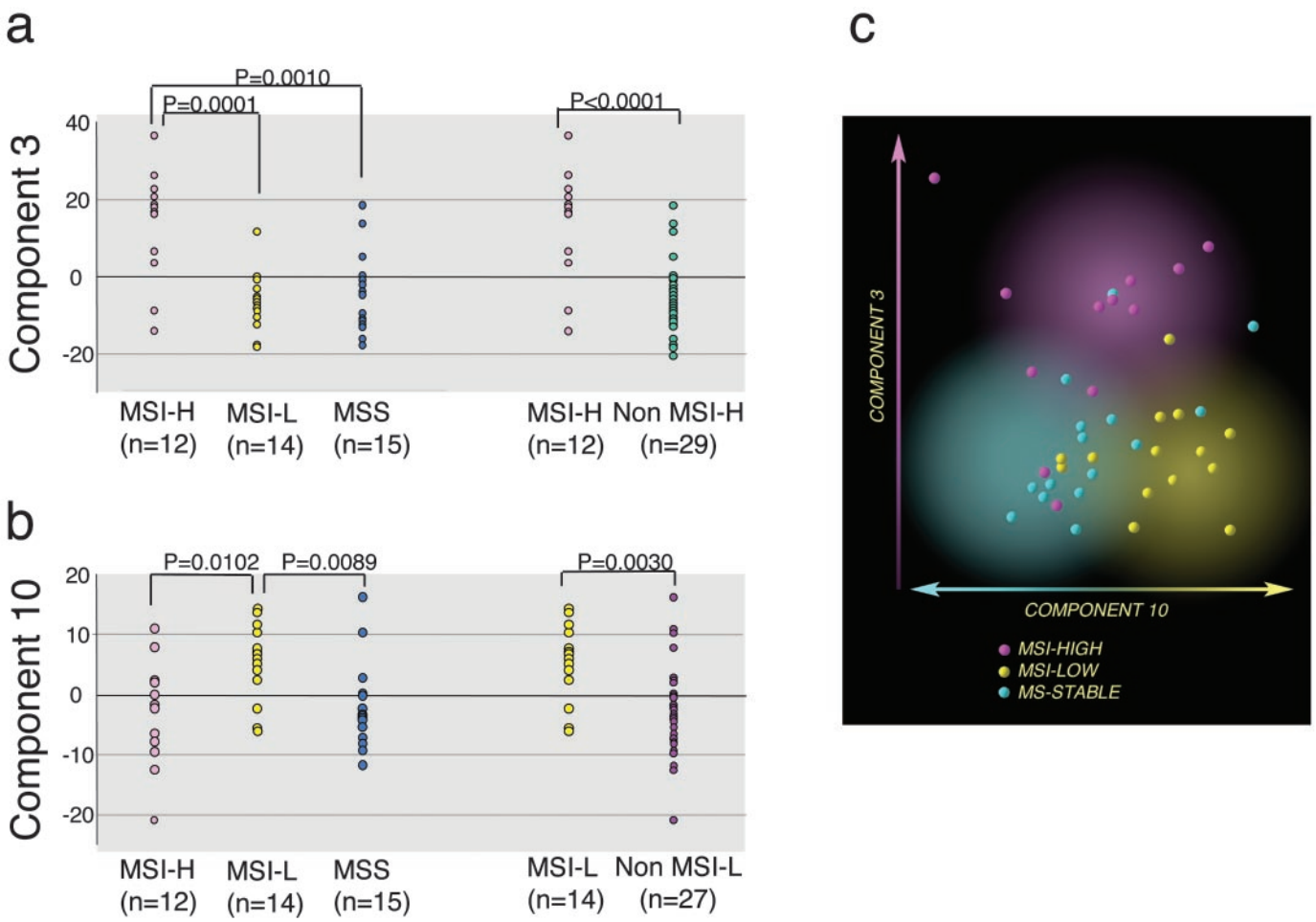

Fig. 1. Scatter plots of output values in components 3 and 10 *versus* MSI status. Informative displays of differences in output values among different MSI groups for components 3 and 10. *a*, one-dimensional scatter plot of discriminative power of component 3. This one-dimensional scatter plot shows that output values for MSI-H tumors were significantly greater than those for MSI-L and MSS tumors (*left three clusters*), as well as those for non-MSI-H tumors (*right two clusters*). *P*s were calculated by student *t* test. *b*, one-dimensional scatter plot of discriminative power of component 10. Output values for MSI-L tumors were significantly greater than those for MSS or MSI-H tumors (*left three clusters*), as well as those for non-MSI-L tumors (*right two clusters*). *P*s were calculated by student *t* test. *c*, two-dimensional scatter plot of values for components 3 (*Y* axis) and 10 (*X* axis). *Violet dots* represent MSI-H tumors, *yellow dots* MSI-L tumors, and *blue dots* MSS tumors. This display suggests that MSI-H, MSI-L, and MSS tumors form three distinct clusters.

relative contributions of each component to the total variance in data were calculated, and components were ranked in decreasing order of their relative contributions. Beginning with the first ranked component, attempts were made to correlate each component with known clinicopathological data. Experimental batch was defined as the source of colon tumor RNAs (Australian *versus* American), and the dates on which RNA extraction and cDNA synthesis were performed (December 2001 *versus* June 2002). The raw data from PCA analysis are also available in a supplemental table on line.[6]

Assessment of Robustness of PCA Results. We used a permutation-based procedure to assess the percentage chances of randomly identifying components segregating the MSI-H and MSI-L categories. Nine hundred ninety-nine permutations consisted of random assignments of MSI status labels (MSI-H, MSI-L, or MSI-S) to the 41 tumors. In each permutation, *P*s were calculated for differences between output values of each component among the three tumor groups (one-way ANOVA). Throughout this permutation procedure, the output value of each component for each tumor was fixed to the value observed in our actual microarray experiments. The number of tumors assigned to each MSI status label was fixed to the true figure (12 for MSI-H, 14 for MSI-L, and 15 for MSS groups). Next, in each permutation, 40 *P*s calculated for each of the 40 components, and the rank and absolute value of each *P* were recorded. After all 999 permutations, *P*s were grouped according to their ranks and compared with *P*s obtained in the actual microarray-PCA analysis. One-way ANOVA calculations were performed in Statistica (StatSoft, Tulsa, OK).

[6] Internet address: http://microarray.umaryland.edu/manuscripts/MoriCAN/.

RESULTS

The Impact of MSI Status and Clinical Features on GMP. Forty-one primary colon cancers were analyzed using cDNA microarrays. To uncover natural groupings based on global gene expression data, unsupervised similarity analyses were used. The first method, hierarchical agglomerative clustering, grouped specimens according to their experimental batch, rather than by molecular or clinical traits (data not shown). Next, we applied PCA, which, in contrast to hierarchical clustering, permits sample classification in multiple independent dimensions (components).

Correlations between each component and molecular or clinical features were tested for statistical significance (Table 1). Component 3 correlated significantly with MSI status (MSI-H, MSI-L, or MSI-S: $P < 0.0001$, one-way ANOVA test; Table 1). This component, designated the MSI-H separator, significantly distinguished MSI-H tumors from non-MSI-H tumors in a two-way comparison ($P < 0.0001$, student *t* test; Table 1 and Fig. 1*a*). Of all components correlating with clinical or molecular features, the MSI-H separator had the highest impact, accounting for 7% of total data variability (Table 1).

Notably, component 10 also correlated significantly with MSI status ($P = 0.0122$, one-way ANOVA test; Table 1), accounting for

Table 2 *A list of genes with the highest loading values in components 3 and 10*

This table displays genes with the twenty highest loading factor values in components 3 and 10. Genes mentioned in the Results and the Discussion sections of the text are indicated by *bold* letters. A positive loading value indicates that for a given component, there is a positive correlation between expression of this gene and the output (e.g., MSI-H, component 3). A negative loading value indicates that there is a negative correlation between expression of this gene and the output for a given component.

Component 3			Component 10		
Loading	Accession no.	Gene name	Loading	Accession no.	Gene name
0.0838	NM_004335	*Bone marrow stromal cell antigen 2*	0.0993	NM_005564	*Lipocalin 2*
0.0774	NM_002909	*Regenerating islet-derived 1 α*	0.0881	NM_001511	***GRO1 oncogene***
0.0741	NM_000582	*Osteopontin*	0.0804	NM_002423	*Matrix metalloproteinase 7*
0.0734	NM_002456	***Mucin 1***	0.0725	AC005020	*Cytochrome P450, subfamily IIIA*
0.0674	XM_039877	***Mucin 5, subtype B***	0.0684	NM_006398	*Ubiquitin D*
0.0600	NM_004184	*Tryptophanyl-tRNA synthetase*	0.0642	NM_014220	*Transmembrane 4 superfamily member 1*
0.0599	NM_006408	*Anterior gradient 2 homologue*	0.0631	NM_000716	*Complement component 4 binding protein, β*
0.0559	M10942	*Metallothionein 1E*	0.0616	NM_005629	*Solute carrier family 6, member 8*
0.0548	NM_002422	*Matrix metalloproteinase 3*	0.0569	NM_003516	*H2A histone family, member O*
0.0527	NM_004223	*Ubiquitin-conjugating enzyme E2L 6*	0.0560	NM_004987	*LIM and senescent cell antigen-like domains 1*
−0.0519	NM_003739	*3-alpha hydroxysteroid dehydrogenase, type II*	0.0549	NM_001165	*Baculoviral IAP repeat-containing 3*
−0.0535	NM_013230	*CD24 antigen*	−0.0549	NM_005410	***Selenoprotein P, plasma, 1***
−0.0564	NM_001482	*Glycine amidinotransferase*	−0.0571	NM_002125	*MHC class II, DR β 5*
−0.0584	NM_003883	***Histone deacetylase 3***	−0.0607	NA	*EST*
−0.0631	NM_006418	*Differentially expressed in hematopoietic lineages*	−0.0627	NM_000597	*Insulin-like growth factor binding protein 2*
−0.0646	NM_007127	*Villin 1*	−0.0649	NM_001846	*Collagen, type IV, α 2*
−0.0669	NM_003944	*Selenium binding protein 1*	−0.0678	NM_000849	*Glutathione S-transferase M3*
−0.0672	NM_138611	*Cervical cancer oncogene 4*	−0.0733	NM_004617	*Transmembrane 4 superfamily member 4*
−0.0733	NM_005518	*3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2*	−0.0948	NM_004413	*Dipeptidase 1 (renal)*
−0.0857	NM_001443	*Fatty acid binding protein 1, liver*	−0.1154	NM_002909	*Regenerating islet-derived 1 α*

2.5% of total data variability. This component, designated the MSI-L separator, differentiated MSI-L tumors from MSI-H and MSS tumors ($P = 0.0030$, student t test; Table 1 and Fig. 1*b*). When components 3 and 10 were combined, the samples tended to form three main groups, according to their MSI status (Fig. 1*c*).

Differentiation grade and anatomic location, which not surprisingly correlated in turn with MSI status (see below), also correlated with component 3 ($P = 0.0014$ and $P = 0.0043$, student t test, respectively; Table 1). Components 1 and 2 correlated with experimental batch, in agreement with hierarchical clustering results (data not shown; see above). Finally, a significant association was also observed between component 6 and clinical stage ($P = 0.0036$, one-way ANOVA test; Table 1), as well as lymph node metastasis ($P = 0.0016$, student t test; Table 1).

Loading value is the number assigned by PCA to represent the influence, within a particular component, of a given gene relative to other genes. Thus, the greater the relative impact of a given gene on a particular component, the greater its loading value (which can be either positive or negative). Genes with the greatest loading values in components 3 and 10 are listed in Table 2. Several of these genes have known links to cancer (see "Discussion"). Numerous additional genes previously associated with MSI-H, MSI-L, or MSS also possessed high loading values in components 3 and 10 (Refs. 10, 18, 22, 42, 43; Table 3).

Statistical Validation of Correlations between Principal Components and MSI Status. To test the robustness of our analysis, we applied a permutation-based procedure that assessed the chances of randomly identifying components segregating the MSI-H and MSI-L categories (Table 4). Permutation analysis is one of several methods useful for the confirmation of statistical correlations. Additional approaches include jackknifing and use of an independent test set (44). Each permutation randomly shuffled the MSI-H, MSI-L, and MSI-S labels for each tumor, then calculated the P of the one-way ANOVA test for the ability of each principal component to separate tumors according to MSI status. Throughout this permutation procedure, the output value at each component for each tumor and the number of tumors assigned to each MSI status label were fixed to the value observed in our actual microarray experiments. Forty Ps, each corresponding to one of the 40 principal components, were then ranked and the lowest P (here referred to as the first-ranked P) was recorded. The above shuffling-ranking procedure was repeated for 999 permutations. In this way, we derived the 999 lowest Ps (the first-ranked Ps) that were obtained purely by chance. Similar procedures were performed on the second-ranked to fortieth-ranked Ps, respectively. A false-finding rate was derived by adding the number of permutations in which the P of a given component ranked higher than it did in the actual data. For example, for component 3, the false-finding rate was only 0.003: *i.e.*, only 3 among 1000 permutations (999 permutations plus the correct MSI labeling) exhibited smaller first-ranked Ps relative to the first-ranked P for the truly labeled microarray data. Similarly, second-ranked Ps lower than that obtained for component 10 occurred in only 64 among the 1000 calculations. Thus, the false-finding rate for component 10 was 0.064.

Correlations between MSI Status and Clinical and Histological Characteristics. MSI status also correlated with histological differentiation and anatomic location ($P < 0.0001$ and $P = 0.0035$, respec-

Table 3 *Loading values for genes known or suspected to be related to MSI status*

Loading values in components 3 and 10 are shown for genes previously reported to have altered expression patterns within the three MSI subtypes. Alterations that have been previously reported in these genes are listed under the "*Mechanism of Alteration*" column.

		Loading Value			
Accession no.	Gene	Component 3	Component 10	Mechanism of alteration	Frequency
NM_000546	*TP53*	0.0108	0.0246	Mutation	MSS, MSI-L > MSI-H
NM_004985	*K-RAS*[a]	−0.0043	0.0015	Mutation	MSS, MSI-L > MSI-H
NM_004324	*BAX*	−0.0152	0.0157	Mutation (causes low mRNA expression)	MSI-H > MSI-L, MSS
NM_000249	*MLH1*	−0.0365	−0.0128	Promoter hypermethylation	MSI-H > MSI-L, MSS
NM_000251	*MSH2*	−0.0009	0.0104	Mutation	MSI-H > MSI-L, MSS
NM_002412	*MGMT*	−0.0301	−0.0114	Promoter hypermethylation	MSI-L > MSS

[a] Average of data from two separate clones on the microarray.

Table 4 *Robustness of the correlation between MSI status and components 3 and 10 studied by a permutation-based significance testing*

This table represents a summary of the permutation-based calculation of the chances of randomly identifying components segregating tumors according to their MSI status. Nine hundred and ninety-nine permutations consisted of random assignments of MSI status labels (MSI-H, -L, or -S) to the 41 specimens. In each permutation, *P*s among the three tumor MSI status groups were calculated (one-way ANOVA), and the 40 components were ranked by *P* from lowest to highest. The *false finding rate* represents the probability of obtaining, purely by chance, the same or a higher level of significance for the correlation between tumor MSI status and PCA component. *Rank*: the rank of a given *P* among 40 *P*s at each permutated or actual MSI-status labeling. The lower the number, the lower the *P*; *false finding rate*: number of *false finding* incidents among the 999 permutations that exhibited *P*s lower than that of the actual microarray experiment/1000.

	Results of the actual microarray experiment		Results from 999 permutations						
Rank	Component	*P*	Median *P*	Minimum *P*	Maximum *P*	5-percentile *P*	95-percentile *P*	No. of false finding	False finding rate
1	3	0.00005	0.01598	0.00003	0.10945	0.00138	0.05330	3	0.003
2	10	0.01218	0.03969	0.00167	0.13045	0.01073	0.08217	64	0.064
3	23	0.08089	0.06630	0.00756	0.16726	0.02912	0.11460	702	0.702
4	6	0.09994	0.09099	0.01486	0.21143	0.04777	0.14443	625	0.625
5	20	0.11606	0.11617	0.02886	0.22414	0.06739	0.17140	497	0.497

tively; Kruskal-Wallis test, Table 5). Specifically, MSI-H tumors were predominantly poorly differentiated and arose more often in the proximal colon, whereas MSI-L and MSS tumors were moderately differentiated and originated equally frequently in the proximal and distal colon ($P < 0.0001$, Mann-Whitney test; $P = 0.0015$, Fisher's exact test, respectively; Table 5). These findings are consistent with previous reports (1, 2). Conversely, neither age, gender, nor clinical stage differed significantly among the three MSI status groups ($P = 0.4152$, $P = 0.1656$ and $P = 0.6212$, respectively; Kruskal-Wallis test; Table 5). Because one of our goals was to define MSI-L tumors, we also studied their association with clinical and histological features. Not surprisingly, there was no significant difference between MSI-L and MSS groups in terms of age ($P = 0.1754$, Mann-Whitney test), gender ($P = 0.1394$, Fisher's test), clinical stage ($P = 0.9355$, Mann-Whitney test), differentiation ($P = 0.9397$ in Mann-Whitney test), or anatomical location ($P = 0.4328$ in Fisher's test; Table 5). This finding is in agreement with previous reports regarding the clinical features of these tumors (20).

DISCUSSION

The aim of the current study was to assess the impact of microsatellite instability status on the GMP of colorectal tumors. In this context, MSI-H colon cancers were revealed by PCA to be distinct from non-MSI-H cancers. This MSI-H separator component exerted a more dominant impact than did any other clinicopathological feature on the GMP of these tumors. These findings support the hypothesis that MSI-H cancers comprise a distinct biologic subgroup (1–6).

In contrast to MSI-H tumors, the MSI-L group of tumors has eluded definition (16, 21). In this context, the current study showed that MSI-L tumors could be distinguished from MSS and MSI-H tumors based on GMP. Moreover, the impact of MSI-L was more subtle than that of MSI-H on molecular phenotype, in agreement with known clinical similarities between MSI-L and MSS cancers (17, 25).

Interestingly, several genes identified on the basis of high loading values in the MSI-H and MSI-L separator components (components 3 and 10, respectively) have been implicated in the biology of MSI-H and MSI-L tumors. For example, relatively high loading values of +0.0734 and +0.0674, respectively, were observed for the cell surface mucin glycoproteins *mucin 1* and *mucin 5, subtype B* in component 3 (Table 2), consistent with frequent overexpression of these genes in MSI-H cancers (45). Similarly, *hMLH1* showed a high negative loading value in component 3 (Table 3), in agreement with this gene's silencing in sporadic MSI-H tumors causing dysfunction in the MMR system (7). *BAX*, an apoptotic regulatory gene, also exerted a negative loading value in component 3 (Table 3), supporting previous reports of its decreased expression in MSI-H tumors (46). The loading value of the DNA repair gene *MGMT* in component 10 (the MSI-L separator) was negative, consistent with decreased expression of *MGMT* in MSI-L tumors due to promoter hypermethylation (22). Furthermore, there were potentially relevant genes with relatively high loading values that had not been previously explored in relation to MSI status. This group of genes included the histone deacetylase gene *HDAC3*, involved in transcriptional regulation through histone-DNA interaction (−0.0584 in component 3; Table 2) and the *GRO1* oncogene, a chemokine ligand gene that stimulates cell proliferation (+0.0881 in component 10; Table 2). Another potentially relevant gene was selenoprotein *P*, an extracellular antioxidant gene (−0.0549 in component 10; Table 2), considering that oxidative stress, already implicated in chronic inflammation, has recently been suggested as a cause of disrupted DNA MMR in MSI-L tumors (47).

In conclusion, the current data show that MSI-H exerts a dominant impact on GMP in colorectal cancers. Finally, these findings support

Table 5 *Summary of clinical features of the colon cancer patients*

The clinical features and MSI status of the tumors are shown. There were no significant differences among MSI-H, -L, and -S tumors in Dukes stage, gender, or age. As expected, there was a preponderance of right-sided and poorly differentiated tumors in the MSI-H category. *NA*, not identified; *MD*, moderately differentiated; *MP*, moderately to poorly differentiated; *PD*, poorly differentiated; *SD*, standard deviation.

	Differentiation				Location			Dukes stage					Gender		Age								
MSI status	MD	MP	PD	NA	R	L	NA	A	B	C	D	NA	Male	Female	<50	50–59	60–69	70–79	≥80	Max	Min	Mean	SD
H ($n = 12$)	2	4	6	0	11	1	0	2	6	4	0	0	4	8	1	2	5	1	3	90	36	67.9	14
L ($n = 14$)	12	1	1	0	6	7	1	0	9	2	3	0	6	8	0	1	5	7	1	88	57	71.3	7.6
S ($n = 15$)	12	2	0	1	4	11	0	1	7	5	1	1	11	4	1	4	5	3	2	91	49	66.1	12
	$P < 0.0001$ (Kruskal-Wallis) H *versus* L + S: $P < 0.0001$ (Mann-Whitney)				$P = 0.0035$ (Kruskal-Wallis) H *versus* L + S: $P = 0.0015$ (Fisher)			$P = 0.6212$ (Kruskal-Wallis)					$P = 0.1656$ (Kruskal-Wallis)		$P = 0.4152$ (Kruskal-Wallis)								

the existence of MSI-L tumors as a distinct molecular genetic category and demonstrate the need for additional research into their molecular origins and clinical significance.

REFERENCES

1. Thibodeau, S. N., Bren, G., and Schaid, D. Microsatellite instability in cancer of the proximal colon. Science (Wash. DC), *260:* 816–819, 1993.
2. Lothe, R. A., Peltomaki, P., Meling, G. I., Aaltonen, L. A., Nystrom-Lahti, M., Pylkkanen, L., Heimdal, K., Andersen, T. I., Moller, P., Rognum, T. O., *et al*. Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history, Cancer Res., *53:* 5849–5852, 1993.
3. Halling, K. C., French, A. J., McDonnell, S. K., Burgart, L. J., Schaid, D. J., Peterson, B. J., Moon-Tasson, L., Mahoney, M. R., Sargent, D. J., O'Connell, M. J., Witzig, T. E., Farr, G. H., Jr., Goldberg, R. M., and Thibodeau, S. N. Microsatellite instability and 8p allelic imbalance in stage B2 and C colorectal cancers. J. Natl. Cancer Inst. (Bethesda), *91:* 1295–303, 1999.
4. Gryfe, R., Kim, H., Hsieh, E. T., Aronson, M. D., Holowaty, E. J., Bull, S. B., Redston, M., and Gallinger, S. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N. Engl. J. Med., *342:* 69–77, 2000.
5. Chao, A., Gilliland, F., Willman, C., Joste, N., Chen, I. M., Stone, N., Ruschulte, J., Viswanatha, D., Duncan, P., Ming, R., Hoffman, R., Foucar, E., and Key, C. Patient and tumor characteristics of colon cancers with microsatellite instability: a population-based study, Cancer Epidemiol. Biomark. Prev., *9:* 539–544, 2000.
6. Wright, C. M., Dent, O. F., Barker, M., Newland, R. C., Chapuis, P. H., Bokey, E. L., Young, J. P., Leggett, B. A., Jass, J. R., and Macdonald, G. A. Prognostic significance of extensive microsatellite instability in sporadic clinicopathological stage C colorectal cancer. Br. J. Surg., *87:* 1197–202, 2000.
7. Cunningham, J. M., Christensen, E. R., Tester, D. J., Kim, C. Y., Roche, P. C., Burgart, L. J., and Thibodeau, S. N. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. Cancer Res., *58:* 3455–3460, 1998.
8. Veigl, M. L., Kasturi, L., Olechnowicz, J., Ma, A. H., Lutterbaugh, J. D., Periyasamy, S., Li, G. M., Drummond, J., Modrich, P. L., Sedwick, W. D., and Markowitz, S. D. Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. Proc. Natl. Acad. Sci. USA, *95:* 8698–702, 1998.
9. Herman, J. G., Umar, A., Polyak, K., Graff, J. R., Ahuja, N., Issa, J. P., Markowitz, S., Willson, J. K., Hamilton, S. R., Kinzler, K. W., Kane, M. F., Kolodner, R. D., Vogelstein, B., Kunkel, T. A., and Baylin, S. B. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. USA, *95:* 6870–6875, 1998.
10. Kane, M. F., Loda, M., Gaida, G. M., Lipman, J., Mishra, R., Goldman, H., Jessup, J. M., and Kolodner, R. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. Cancer Res., *57:* 808–811, 1997.
11. Leach, F. S., Nicolaides, N. C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomaki, P., Sistonen, P., Aaltonen, L. A., Nystrom-Lahti, M., *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell, *75:* 1215–1225, 1993.
12. Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., Kane, M., Earabino, C., Lipford, J., Lindblom, A., *et al*. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature (Lond.), *368:* 258–261, 1994.
13. Papadopoulos, N., Nicolaides, N. C., Wei, Y. F., Ruben, S. M., Carter, K. C., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., Fraser, C. M., Adams, M. D., *et al*. Mutation of a mutL homolog in hereditary colon cancer. Science (Wash. DC), *263:* 1625–1629, 1994.
14. Lengauer, C., Kinzler, K. W., and Vogelstein, B. Genetic instability in colorectal cancers. Nature (Lond.), *386:* 623–627, 1997.
15. Perucho, M. Microsatellite instability: the mutator that mutates the other mutator. Nat. Med., *2:* 630–631, 1996.
16. Jass, J. R., Biden, K. G., Cummings, M. C., Simms, L. A., Walsh, M., Schoch, E., Meltzer, S. J., Wright, C., Searle, J., Young, J., and Leggett, B. A. Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. J. Clin. Pathol. (Lond.), *52:* 455–460, 1999.
17. Gafa, R., Maestri, I., Matteuzzi, M., Santini, A., Ferretti, S., Cavazzini, L., and Lanza, G. Sporadic colorectal adenocarcinomas with high-frequency microsatellite instability. Cancer (Phila.), *89:* 2025–2037, 2000.
18. Samowitz, W. S., Holden, J. A., Curtin, K., Edwards, S. L., Walker, A. R., Lin, H. A., Robertson, M. A., Nichols, M. F., Gruenthal, K. M., Lynch, B. J., Leppert, M. F., and Slattery, M. L. Inverse relationship between microsatellite instability and *K-ras* and *p53* gene alterations in colon cancer. Am. J. Pathol., *158:* 1517–1524, 2001.
19. Boland, C. R., Thibodeau, S. N., Hamilton, S. R., Sidransky, D., Eshleman, J. R., Burt, R. W., Meltzer, S. J., Rodriguez-Bigas, M. A., Fodde, R., Ranzani, G. N., and Srivastava, S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res., *58:* 5248–5257, 1998.
20. Tomlinson, I., Halford, S., Aaltonen, L., Hawkins, N., and Ward, R. Does MSI-low exist? J. Pathol., *197:* 6–13, 2002.
21. Jass, J. R., Young, J., and Leggett, B. A. Biological significance of microsatellite instability-low (MSI-L) status in colorectal tumors. Am. J. Pathol., *158:* 779–781, 2001.
22. Whitehall, V. L., Walsh, M. D., Young, J., Leggett, B. A., and Jass, J. R. Methylation of *O*-6-methylguanine DNA methyltransferase characterizes a subset of colorectal cancer with low-level DNA microsatellite instability. Cancer Res., *61:* 827–830, 2001.
23. Laiho, P., Launonen, V., Lahermo, P., Esteller, M., Guo, M., Herman, J. G., Mecklin, J. P., Jarvinen, H., Sistonen, P., Kim, K. M., Shibata, D., Houlston, R. S., and Aaltonen, L. A. Low-level microsatellite instability in most colorectal carcinomas. Cancer Res., *62:* 1166–1170, 2002.
24. Halford, S., Sasieni, P., Rowan, A., Wasan, H., Bodmer, W., Talbot, I., Hawkins, N., Ward, R., and Tomlinson, I. Low-level microsatellite instability occurs in most colorectal cancers and is a nonrandomly distributed quantitative trait. Cancer Res., *62:* 53–57, 2002.
25. Thibodeau, S. N., French, A. J., Cunningham, J. M., Tester, D., Burgart, L. J., Roche, P. C., McDonnell, S. K., Schaid, D. J., Vockley, C. W., Michels, V. V., Farr, G. H., Jr., and O'Connell, M. J. Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1. Cancer Res., *58:* 1713–1718, 1998.
26. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, *95:* 14863–8, 1998.
27. Selaru, F. M., Zou, T., Xu, Y., Shustova, V., Yin, J., Mori, Y., Sato, F., Wang, S., Olaru, A., Shibata, D., Greenwald, B. D., Krasna, M. J., Abraham, J. M., and Meltzer, S. J. Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays. Oncogene, *21:* 475–478, 2002.
28. Xu, Y., Selaru, F. M., Yin, J., Zou, T. T., Shustova, V., Mori, Y., Sato, F., Liu, T. C., Olaru, A., Wang, S., Kimos, M. C., Perry, K., Desai, K., Greenwald, B. D., Krasna, M. J., Shibata, D., Abraham, J. M., and Meltzer, S. J. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res., *62:* 3493–3497, 2002.
29. Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput., 455–66, 2000.
30. Kachigan, S. K. Multivariate statistical analysis: a conceptual introduction, p. xii, 297. New York: Radius Press, 1982.
31. Chomczynski, P., and Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem., *162:* 156–159, 1987.
32. Meltzer, S. J., Yin, J., Manin, B., Rhyu, M. G., Cottrell, J., Hudson, E., Redd, J. L., Krasna, M. J., Abraham, J. M., and Reid, B. J. Microsatellite instability occurs frequently and in both diploid and aneuploid cell populations of Barrett's-associated esophageal adenocarcinomas. Cancer Res., *54:* 3379–3382, 1994.
33. Mori, Y., Sato, F., Selaru, F. M., Olaru, A., Perry, K., Kimos, M. C., Tamura, G., Matsubara, N., Wang, S., Xu, Y., Yin, J., Zou, T. T., Leggett, B., Young, J., Nukiwa, T., Stine, O. C., Abraham, J. M., Shibata, D., and Meltzer, S. J. Instabilotyping reveals unique mutational spectra in microsatellite-unstable gastric cancers. Cancer Res., *62:* 3641–3645, 2002.
34. Luo, L., Salunga, R. C., Guo, H., Bittner, A., Joy, K. C., Galindo, J. E., Xiao, H., Rogers, K. E., Wan, J. S., Jackson, M. R., and Erlander, M. G. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nat. Med., *5:* 117–122, 1999.
35. Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D., and Eberwine, J. H. Amplified RNA synthesized from limited quantities of heterogeneous cDNA, Proc. Natl. Acad. Sci. USA, *87:* 1663–1667, 1990.
36. Alizadeh, A., Eisen, M., Botstein, D., Brown, P. O., and Staudt, L. M. Probing lymphocyte biology by genomic-scale gene expression analysis. J. Clin. Immunol., *18:* 373–379, 1998.
37. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. The transcriptional program in the response of human fibroblasts to serum. Science (Wash. DC), *283:* 83–87, 1999.
38. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA, *96:* 9212–9217, 1999.
39. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.
40. Cleveland, W. S. Lowess: robust locally weighted regression for smoothing and graphing data in two or more dimensions. Graphical techniques for exploring data. Association for Computing Machinery SIGGRAPH Tutorial 23, 1983.
41. Eisen, M. B., and Brown, P. O. DNA arrays for analysis of gene expression. Methods Enzymol., *303:* 179–205, 1999.
42. Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J. C., and Perucho, M. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science (Wash. DC), *275:* 967–969, 1997.
43. Jass, J. R., Whitehall, V. L., Young, J., and Leggett, B. A. Emerging concepts in colorectal neoplasia. Gastroenterology, *123:* 862–876, 2002.
44. Efron, B. Bootstrap methods: another look at the jackknife. Ann. Stat., *7:* 1–26, 1979.
45. Biemer-Huttmann, A. E., Walsh, M. D., McGuckin, M. A., Simms, L. A., Young, J., Leggett, B. A., and Jass, J. R. Mucin core protein expression in colorectal cancers with high levels of microsatellite instability indicates a novel pathway of morphogenesis. Clin. Cancer Res., *6:* 1909–1916, 2000.
46. Ionov, Y., Yamamoto, H., Krajewski, S., Reed, J. C., and Perucho, M. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution, Proc. Natl. Acad. Sci. USA, *97:* 10872–7, 2000.
47. Chang, C. L., Marra, G., Chauhan, D. P., Ha, H. T., Chang, D. K., Ricciardiello, L., Randolph, A., Carethers, J. M, and Boland, C. R. Oxidative stress inactivates the human DNA mismatch repair system. Am. J. Physiol. Cell Physiol., *283:* C148–C154, 2002.