Background Mutation Frequency in Microsatellite-Unstable Colorectal Cancer

Heli Sammalkorpi,[1] Pia Alhopuro,[1] Rainer Lehtonen,[1] Jarno Tuimala,[3] Jukka-Pekka Mecklin,[4] Heikki J. Järvinen,[2] Josef Jiricny,[5] Auli Karhu,[1] and Lauri A. Aaltonen[1]

[1]Department of Medical Genetics, Biomedicum Helsinki, University of Helsinki; [2]Second Department of Surgery, Helsinki University Central Hospital, Helsinki, Finland; [3]CSC, Finnish IT Center for Science, Espoo, Finland; [4]Department of Surgery, Jyväskylä Central Hospital, Jyväskylä, Finland; and [5]Institute of Molecular Cancer Research, University of Zurich, Zurich, Switzerland

Abstract

Microsatellite instability (MSI) is observed in ~12% of colorectal cancers. Genes containing a mononucleotide microsatellite in the coding sequence are particularly prone to inactivation in MSI tumorigenesis, and much work has been conducted to identify genes with high repetitive tract mutation rates in these tumors. Much less attention has been paid to background mutation frequencies, and no work has focused on nontranscribed regions. Here, we studied 114 nontranscribed intergenic A/T and C/G repeats 6 to 10 bp in length, located distant from known genes, to examine background mutation frequencies in MSI colorectal cancers. A strong correlation with tract length was observed, and mutation frequencies of up to 87% were observed in 8 to 10 bp tracts. Subsequently, to compare the background mutation rate in transcribed and nontranscribed noncoding repeats, we screened nine randomly selected intronic C/G8 repeats. In addition, the coding repeats of seven suggested MSI target genes, and nine previously published intronic A8 and G8 repeats were analyzed. Intronic repeats seemed to mutate less frequently than nontranscribed intergenic repeats. Our results show that strand slippage mutations in mismatch repair–deficient cells are as abundant in short intergenic repeats as in many proposed MSI target genes. However, under mismatch repair deficiency, strand slippage mutations in transcribed sequences seem to be repaired more efficiently than in intergenic nontranscribed sequences. The mechanisms causing these differences are not yet understood and should be a subject for further studies. For MSI target gene identification, repeats in transcribed sequences seem to be the most appropriate reference group for coding region repeat mutations. [Cancer Res 2007;67(12):5691–8]

Introduction

Microsatellite instability (MSI) manifests as small deletions and insertions in short repetitive sequences genome-wide, and is caused by a defective DNA mismatch repair (MMR) system. MSI is detected in 10% to 15% of colorectal, endometrial, and gastric cancers (1, 2), and is associated with hereditary nonpolyposis colorectal cancer syndrome (3–6). The evolution of MSI tumors is a continuous process of mutations and selection favoring neoplastic growth, and the driving force for carcinogenesis is mutations in MSI target genes. In 1997, the National Cancer Institute workshop set criteria to distinguish real MSI target genes from bystander mutation targets. These included (*a*) a high frequency of mutations, (*b*) biallelic inactivation, (*c*) involvement in a growth suppressor pathway, (*d*) inactivation of the same pathway in MSS tumors, and (*e*) functional studies in *in vitro* or *in vivo* models (2). More recently, statistical tools for identifying real MSI target genes have been developed (7, 8).

The cornerstone of target gene identification has been high mutation rate. The first proposed MSI target genes were *TGFβRII* (containing A10 repeat), *BAX* (G8), and *IGFIIR* (G8) showing mutation frequencies of 82%, 51%, and 9%, respectively, in MSI colorectal cancer (CRC) tumors (9–11). Since then, tens of novel MSI target genes have been proposed (8), but few functional studies have provided evidence for a causative role in MSI carcinogenesis (12, 13).

In MMR-deficient cells, mutations in noncoding repeats distant from splice sites are unlikely to promote carcinogenesis. Background mutation rates in MSI cancers have been studied only in intronic repeats probably because full intergenic sequences were not available at the time, and one would not expect a fundamental difference between strand slippage mutations in intronic and intergenic mononucleotide tracts under MMR deficiency. The detected intronic mutation rates have been low: studies of 18 randomly chosen intronic repeats length 6 to 10 bp, screened as control repeats, show somatic mutation rates of ≤6% in MSI CRC tumors (8, 14). In a study of 29 intronic A8 and G8 repeats by Zhang et al., nine repeats showed high mutation frequencies between 25% and 54% in CRC cell lines (15), but some of the repeats were not examined in germ line tissue. Two other studies have examined identical tracts as Zhang et al. but with primary CRC tumors and corresponding normal tissues (16, 17), and in these studies, high mutation rates were not confirmed. Our recent pilot study of 10 intergenic A/T9 microsatellites showed high somatic mutation rates in MSI CRCs, 70% being the highest detected (18). We wished to extend the study to examine whether high mutation rates were indeed commonly detected in intergenic sequences, and whether mutation rates in transcribed regions—evaluated with the same material and methodology—would indeed display lower mutation rates.

In this study, we report the analyses of somatic background mutation rates in 114 short intergenic nontranscribed A/T and C/G mononucleotide repeats in MSI CRC. Additionally, to compare the mutation frequencies of intergenic repeats with neutral intronic transcribed repeats, we sequenced nine randomly chosen intronic C/G8 repeats. To examine whether the sensitivity of our methodology played a role in the results, we also examined mutations in seven previously proposed MSI target genes and nine previously published intronic repeats.

Requests for reprints: Lauri A. Aaltonen, Department of Medical Genetics, Biomedicum Helsinki, University of Helsinki, P.O. Box 63, FIN-00014 Helsinki, Finland. Phone: 358-91912-5595; Fax: 358-91912-5105; E-mail: lauri.aaltonen@helsinki.fi.

doi:10.1158/0008-5472.CAN-06-4314

Materials and Methods

Sample selection. The CRC samples and corresponding normal tissues used in this study were selected from a population-based sample set of 1,042 colorectal tumors collected since 1994 (19, 20). The studies were approved by the Helsinki University Hospital Ethics Committee. The sample set for mutation screening of intergenic and intronic regions, and seven coding repeats (*TGFβRII, BLM, CtIP, MSH3, MSH6, IGFIIR*, and *BAX*) consisted of 30 MSI-High colorectal adenocarcinomas; 20 sporadic tumors and 10 cases carrying a germ line *MLH1* or *MSH2* mutation. For intergenic repeats AC105204_2, AC105411, AC104013, AL442644, and AC093511, and for the seven coding repeats, a larger sample panel of an additional 70 MSI-High CRCs was also used, consisting of 54 sporadic tumors and 16 cases with a germ line *MLH1* or *MSH2* mutation. A sample panel of *TCF-4* sequencing has been described in our previous study (18). Genomic DNA was extracted from fresh-frozen specimens evaluated by a pathologist, and 90% of the samples displayed >60% carcinoma tissue. The normal tissue DNA was extracted from blood or normal colonic epithelium distant from the site of the tumor.

Intergenic and intronic DNA sequences. One hundred and forty-one intergenic mononucleotide repeats were selected randomly, distant from known genes, from human chromosomes 1 to 22 using the Ensembl database (release 28, February 2005).[6] Telomeric and centromeric areas, and X and Y chromosomes were excluded. Fifty-one percent of the repeats were located >1 Mb from known genes, and 91% were located ≥0.1 Mb from known genes. AC105204_2, AC104013, and AL442644 were located >1 Mb whereas AC105411 and AC093511 were located 0.2 Mb from known genes, respectively. The selection of nine previously published intronic repeats is described in Zhang et al. (15), and the nine additional intronic C8 and G8 repeats were selected randomly from chromosomes 7 to 16 using the Ensembl database (release 40, August 2006).[6] Some of the loci were polymorphic. Data was not gathered from experiments in which germ line DNA was found to display other sequences than the expected wild-type allele in homozygous form.

Mutation screening. Primers were designed by using the Primer3 program[7] (primer sequences and PCR conditions are available on request). All the fragments with an 8 to 10 bp repeat were amplified using proof-reading enzyme Phusion (Finnzymes). PCR products were purified enzymatically using ExoSAP-IT reagent (U.S. Biochemical Corporation) according to the manufacturer's instructions. Direct sequencing was done by using Big Dye Terminator kit 3.1 (Applied Biosystems), and ABI3730 Automatic DNA Sequencer (Applied Biosystems) according to the manufacturer's instructions. If a mutation was detected in a tumor, corresponding normal tissue was always analyzed to confirm their somatic origins. Mutation signals >10% were scored as mutations (compared with wild-type signals). If the signal of the mutant allele was 10% to 30% of the corresponding wild-type signal, particular attention was paid to the tumor percentage of that specific tumor to differentiate between nonclonal and clonal mutations. *TCF-4* sequencing has been described in our previous study (18). In three intergenic repeats (AF015262, AP004835, and AL442183), reference sequences from the database showed incorrect repeat lengths. In these cases, repeats were analyzed according to the true genotype detected in our sample set.

Flanking sequence. Flanking sequences of the 114 intergenic repeats were extracted from Ensembl database (release 28, February 2005).[6] C+G content, and evolutionary conservation were analyzed for surrounding sequences of the 114 repeats. C+G content was determined 500 bp upstream and downstream from the repeat. The VISTA alignment tool[8] was used to analyze the conservation of flanking sequences between human and other species (mouse, rat, fugu, chicken, frog, dog, cow, and opossum) 2,500 bp on both 5′ and 3′ directions from the repeat. Probability threshold $P = 0.5$, calculation windows 25 and 100 bp, and minimum conservation identity of 70% were used. In addition, from each repeat type 7 to 10 bp in length, we selected two repeats with the highest mutation rates ($n = 16$). The flanking sequences of these repeats were searched for possible shared DNA motifs. To identify DNA motifs 500 bp upstream and downstream from the repeats, we used tools from Genomatix software package (Genomatix Suite, release 3.4.1, including GEMS Launcher software package, release 4.2.1). First, we used CoreSearch tool (21) to detect ≤20 bp DNA sequences (motifs) located in the flanking sequence of the selected repeats ($n = 16$, threshold 0.9), and identified six shared motifs. Then, we used MatInspector tool (22) to identify the prevalence of these six detected motifs in the flanking sequences of all succeeding repeats ($n = 114$). More detailed information on the software is available at Genomatix.[9]

Statistical methods. Before statistical analysis, Kolmogorov-Smirnov test was used to test data normality. Because data was not normally distributed, Spearman correlation (two-tailed) coefficient (r) was used to estimate correlation between mutation rate and repeat length, and C+G content of flanking sequence. Kolmogorov-Smirnov test, and χ^2 test were used to define the distribution of a motif in the data set. χ^2 test was used for other analyses. For all tests, $P < 0.05$ was considered significant. SPSS 12.0.1 (SPSS Inc.) was used for statistical testing.

Results

Screening intergenic repeats. Of 141 intergenic repeats, 114 (81%) were successfully analyzed in at least 83% of the 30 tumor samples (Table 1). The sample panel of 30 MSI CRCs contained no single tumor with significantly more mutations than others; the observed mutation rates within a tumor were normally distributed ($P = 0.21$, Kolmogorov-Smirnov test). Repeats of 6 to 7 bp displayed low mutation rates (average mutation rates, 0.55%, and 3%, respectively; Table 2). Much higher rates were observed in the repeats of 8 to 10 bp, and many of them (25 of 60, 42%) showed a rate of ≥30% (Table 2). In the 8-bp repeats, the highest mutation rate was 37% (average, 13%), in the 9-bp repeats, 81% (average, 32%), and in the 10-bp repeats 92% (average, 50%; Table 2). Repeats with the highest mutation rates were examined in an additional 70 cancers, and as detailed below, the mutation rates in the extended set were typically somewhat lower (Tables 1 and 3). There was a strong positive correlation between repeat length and mutation rate (correlation, 0.81; $P \leq 0.001$, $n = 114$; Fig. 1), both in A/T repeats (correlation, 0.81; $P \leq 0.001$, $n = 69$) and in C/G repeats (correlation, 0.84; $P \leq 0.001$, $n = 45$, Spearman correlation). C/G repeats were more unstable than A/T repeats in 8 to 10 bp lengths ($n = 60$, $P \leq 0.001$, χ^2 test; Table 2). The great majority (97%) of the mutations were heterozygous. Deletions covered 89% of all mutations, whereas insertions were observed in 11% of mutations detected. In 94% of mutant cases, mutant allele peak intensities were >30% of the corresponding wild-type signal. Of all intergenic repeats, 23 showed germ line polymorphisms (Table 1). Polymorphic 8 to 10 bp long repeats typically contained a large variety of heterozygous and homozygous repeat insertions and deletions of different sizes in germ line tissue. Because the amount of length polymorphism was highly variable, in long repeats, the amount of data gathered was in some cases low (Table 1).

Flanking sequence. The flanking sequences of 114 intergenic repeats was analyzed, but no significant correlation between mutation rate and C+G content, or for rate of evolutionarily conserved sequences, was detected. In two repeats with the highest

[6] http://www.ensembl.org/
[7] http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
[8] http://pipeline.lbl.gov/cgi-bin/gateway2
[9] http://www.genomatix.de/

Table 1. Primer locations and somatic mutation rates in 114 intergenic repeats with 30 MSI CRCs and with 100 MSI CRCs

Fragment	Primer location	Repeat	Mutation rate, no. of cases (%)	Germ line polymorphism, no. of detected cases
A/T6 (n = 18)				
AC015771	Chr11:24,288,157–24,288,451	A/T6	7 (2/30)	None
AP001801	Chr11q14:80,835,436–80,835,611	A/T6	3 (1/30)	None
AC079239	Chr4p14:36,359,550–36,359,864	A/T6	0 (0/30)	ND
AL359173_2	Chr6q24:140,729,004–140,729,244	A/T6	0 (0/28)	ND
AL591304	Chr6q16:93,191,308–93,191,602	A/T6	0 (0/28)	ND
AC012238	Chr8q23:116,178,686–116,178,924	A/T6	0 (0/30)	ND
AC067836	Chr8q24:123,100,819–123,100,995	A/T6	0 (0/30)	ND
AL592227	Chr9p23:11,601,021–11,601,223	A/T6	0 (0/30)	ND
AL358214	Chr10p14:9,946,103–9,946,282	A/T6	0 (0/30)	ND
AL021877	Chr22q12:33,264,992–33,265,203	A/T6	0 (0/30)	ND
AC079239_2	Chr4p14:36,359,550–36,359,864	A/T6	0 (0/29)	ND
AL591304_2	Chr6q16:93,191,308–93,191,602	A/T6	0 (0/29)	ND
AC009490	Chr2p16:53,221,891–53,222,179	A/T6	0 (0/29)	ND
AC079239_3	Chr4p14:36,449,781–36,450,040	A/T6	0 (0/28)	ND
AP003532	Chr11q24:127,130,142–127,130,425	A/T6	0 (0/28)	ND
AL136375	Chr1q31:188,042,481–188,042,714	A/T6	0 (0/30)	ND
AC093511	Chr16p12:26,144,741–26,144,981	A/T6	0 (0/29)	ND
AC020637_2	Chr12q14:57,403,851–57,404,145	A/T6	0 (0/30)	ND
C/G6 (n = 11)				
AC110785	Chr4q13:61,217,217–61,217,514	C/G6	3 (1/30)	None
AC006007	Chr7q35:144,799,210–144,799,473	C/G6	3 (1/30)	None
AC024697_2	Chr2q14:116,947,138–116,947,378	C/G6	0 (0/30)	ND
AC009490_2	Chr2p16:53,279,671–53,279,860	C/G6	0 (0/29)	ND
AC009490_3	Chr2p16:53,242,746–53,242,954	C/G6	0 (0/27)	ND
AC079239_4	Chr4p14:36,509,930–36,509,168	C/G6	0 (0/29)	ND
AC004415	Chr7p21:10,505,373–10,505,563	C/G6	0 (0/30)	ND
AC090502	Chr12q21:72,798,886–72,799,126	C/G6	0 (0/28)	ND
AC011459	Chr19q12:36,952,154–36,952,390	C/G6	0 (0/30)	ND
AL499605	Chr1p21:106,876,633–106,876,832	C/G6	0 (0/30)	ND
AC104035	Chr15q26:89,734,786–89,738,108	C/G6	0 (0/30)	ND
A/T7 (n = 14)				
AC024697	Chr2q14:116,853,898–116,854,087	A/T7	10 (3/30)	None
AC091154	Chr17q22:48,457,931–48,458,256	A/T7	10 (3/29)	None
AL445683	Chr9q33:119,531,254–119,531,492	A/T7	8 (2/25)	None
AL590866	Chr6p22:23,347,778–23,347,950	A/T7	7 (2/30)	None
AC024697_3	Chr2q14:116,884,593–116,884,833	A/T7	3 (1/30)	None
AL359173	Chr6q24:140,780,826–140,781,089	A/T7	3 (1/30)	None
AL133229	Chr20q12:39,851,219–39,851,400	A/T7	3 (1/30)	None
AC009490_4	Chr2p16:53,221,891–53,222,179	A/T7	0 (0/29)	ND
AC087798	Chr12q21:83,197,550–83,197,790	A/T7	0 (0/30)	ND
AC008686	Chr19p13:13,689,370–13,689,561	A/T7	0 (0/30)	ND
AC106860	Chr4q32:162,272,865–162,273,085	A/T7	0 (0/30)	ND
AC105204	Chr15q25:85,587,138–85,587,452	A/T7	0 (0/30)	ND
AL360010	Chr1q43:234,907,733–234,907,973	A/T7	0 (0/30)	ND
AL133229_2	Chr20q12:39,851,219–39,851,400	A/T7	0 (0/30)	ND
C/G7 (n = 11)				
AC006041	Ch7p21:15,581,240–15,581,430	C/G7	10 (3/30)	None
AL132719	Chr14q32:98,065,314–98,065,625	C/G7	8 (2/26)	None
AL357353	Ch10q21:57,243,391–57,243,660	C/G7	7 (2/30)	None
AC125492*	Ch12q14:59,517,777–59,517,993	C/G7	4 (1/28)	2/30
AC023793	Ch3q13:119,373,187–119,373,399	C/G7	3 (1/30)	None
AP000946	Ch21q21:20,289,420–20,289,660	C/G7	3 (1/30)	None
AC079239_5	Chr4p14:36,449,781–36,450,040	C/G7	0 (0/28)	ND
AL159995	Ch9p21:30,214,394–30,214,585	C/G7	0 (0/30)	ND
AL137004	Chr6p22:21,378,647–21,372,282	C/G7	0 (0/30)	ND
AC009432	Chr15q26:92,868,231–92,868,547	C/G7	0 (0/27)	ND
AL133229_3	Chr20q12:39,851,219–39,851,400	C/G7	0 (0/30)	ND

(*Continued on the following page*)

Table 1. Primer locations and somatic mutation rates in 114 intergenic repeats with 30 MSI CRCs and with 100 MSI CRCs (Cont'd)

Fragment	Primer location	Repeat	Mutation rate, no. of cases (%)	Germ line polymorphism, no. of detected cases
A/T8 (n = 13)				
AC105204_2 †	Chr15q25:85,554,726–85,555,015	A/T8	21 (6/28); 19 (19/98) †	ND
AC011403	Chr5q34:164,672,247–164,672,470	A/T8	17 (5/30)	None
AL590078	Chr9p21:26,458,917–26,459,149	A/T8	11 (3/28)	None
AL512446	Chr6q22:121,262,503–121,262,692	A/T8	7 (2/30)	None
AL020994	Chr22q12:26,049,315–26,049,515	A/T8	7 (2/28)	None
AC069220	Chr3q26:170,505,818–170,506,052	A/T8	4 (1/26)	None
AC114278	Chr5q21:105,700,434–105,700,668	A/T8	3 (1/30)	None
AL607082	Chr10q25:110,756,404–110,756,579	A/T8	3 (1/30)	None
AL160391	Chr13q21:69,619,386–69,619,622	A/T8	3 (1/30)	None
AL353620*	Chr13q33:103,195,991–103,196,221	A/T8	0 (0/21)	9/30
AL360010_2	Chr1q43:234,907,733–234,907,973	A/T8	0 (0/30)	ND
AL109762	Chr21q21:16,320,188–16,320,378	A/T8	0 (0/28)	ND
AC005939	Chr17q24:65,916,252–65,916,576	A/T8	0 (0/26)	ND
C/G8 (n = 9)				
AC105411 †	Chr16q23:78,968,987–78,969,170	C/G8	37 (11/30); 30 (29/96) †	ND
AL590304	Chr13q21:55,610,375–55,610,640	C/G8	30 (9/30)	None
AL109953	Chr20p11:20,731,397–20,731,623	C/G8	29 (8/28)	None
AC024382*	Chr8q24:127,810,968–127,811,201	C/G8	24 (6/25)	1/26
AC004926	Chr7q21:85,315,286–85,315,477	C/G8	23 (7/30)	None
AF015262*	Chr21q22:35,404,815–35,405,044	C/G8	21 (3/14)	14/28
AC092566	Chr19p13:8,658,095–8,645,8285	C/G8	20 (6/30)	None
AC093252*	Chr5p14:27,893,387–27,893,144	C/G8	20 (4/20)	8/28
AC026473*	Chr16q21:59,678,144–59,678,379	C/G8	10 (3/29)	1/30
A/T9 (n = 10)				
AC104013 †	Chr8q21:84,781,882–84,782,072	A/T9	53 (16/30); 41 (40/98) †	None
AP003532_2	Chr11q24:127,130,142–127,130,425	A/T9	46 (13/28)	None
AL359238	Chr14q31:82,491,728–82,492,034	A/T9	44 (11/25)	None
AL355099	Chr14q21:42,504,619–42,504,958	A/T9	20 (6/30)	None
AL451146*	Chr10q23:85,132,147–85,132,387	A/T9	17 (4/24)	5/29
AL390295	Chr13q13:34,252,765–34,252,999	A/T9	14 (4/29)	None
AL354682	Chr9q21:80,833,540–80,833,778	A/T9	10 (3/30)	None
AC084308	Chr8p12:36,560,351–36,560,536	A/T9	7 (2/27)	None
AC079893	Chr7q31:109,263,477–109,263,656	A/T9	3 (1/30)	None
AC104686	Chr4p15:11,532,088–11,532,308	A/T9	3 (1/30)	None
C/G9 (n = 6)				
AC104209*	Chr18q12:36,068,830–36,069,148	C/G9	81 (17/21)	4/25
AC122718*	Chr5q12:60,785,390–60,785,649	C/G9	70 (16/23)	4/27
AL954650*	Chr1q31:188,658,391–188,658,626	C/G9	63 (17/27)	2/30
AL008725*	Chr20q13:42,929,465–42,929,785	C/G9	29 (2/7)	17/29
AL442183*	Chr14q21:42,955,991–42,956,328	C/G9	28 (5/18)	9/27
AC093412*	Chr3p24:28,800,867–28,801,057	C/G9	23 (4/17)	11/28
A/T10 (n = 14)				
AC093511	Chr16p12:26,144,741–26,144,981	A/T10	76 (22/29); 74 (67/90) †	None
AL355154	Chr13q31:80,916,438–80,916,646	A/T10	67 (18/27)	None
AP004835*	Chr11q22:96,805,039–96,805,259	A/T10	60 (9/15)	13/28
AC011942	Chr11q12:38,307,678–38,307,905	A/T10	50 (15/30)	None
AL163153*	Chr14p21:42,518,318–42,518,596	A/T10	48 (13/27)	3/30
AL035665	Chr20q12:38,702,819–38,703,007	A/T10	41 (12/29)	None
AC079610	Chr2q34:213,922,598–213,922,831	A/T10	37 (11/30)	None
AL592428	Chr6q16:92,711,813–92,712,030	A/T10	35 (9/26)	None
AC087774*	Chr12p12:17,871,052–17,871,286	A/T10	33 (1/3)	23/26
AL357632	Chr1p31:81,375,127–81,375,339	A/T10	30 (9/30)	None
AL163153_2	Chr14q21:42,504,619–42,504,958	A/T10	27 (7/26)	None
AL136375_2*	Chr1q31:188,042,481–188,042,714	A/T10	23 (5/22)	7/29
AC097633*	Chr3p24:30,065,841–30,066,009	A/T10	21 (3/14)	13/27
AP003174	Chr11q23:114,963,073–114,963,258	A/T10	8 (2/26)	None

(*Continued on the following page*)

Table 1. Primer locations and somatic mutation rates in 114 intergenic repeats with 30 MSI CRCs and with 100 MSI CRCs (Cont'd)

Fragment	Primer location	Repeat	Mutation rate, no. of cases (%)	Germ line polymorphism, no. of detected cases
C/G10 (n = 8)				
AL442644 †	Chr9q31:102,316,494–102,316,716	C/G10	92 (24/26); 77 (66/86) †	None
AL353072*	Chr1q31:191,467,569–191,467,753	C/G10	87 (8/9)	10/26
U95743*	Chr16p13:13,782,189–13,782,383	C/G10	87 (13/15)	11/26
AL049833*	Chr14q32:96,621,389–96,621,134	C/G10	75 (6/8)	20/28
AC020637*	Chr12q14:57,403,851–57,404,145	C/G10	69 (18/26)	2/28
AL442126*	Chr13q31:89,538,308–89,538,528	C/G10	67 (4/6)	19/25
AL035456	Chr20p12:10,647,064–10,647,263	C/G10	56 (14/25)	None
AL162852	Chr13q22:77,538,008–77,538,223	C/G10	0 (0/30)	ND

Abbreviation: ND, not determined.
*Repeats with germ line length polymorphisms. All the samples with germ line length polymorphisms were excluded from the analysis.
† The repeats AC105204_2, AC105411, AC104013, AC093511, and AL442644 were analyzed with 100 MSI CRCs.

mutation rates in each group of 7 to 10 bp (n = 16), we found no shared short DNA motifs correlating with high mutation rate.

Screening intergenic and coding repeats with 100 MSI CRCs. Five intergenic repeats with high mutation rates were screened with an extended sample panel of 100 MSI CRCs, and somatic mutation frequencies were consistent with the rates observed in 30 MSI CRCs. Detected mutation frequencies were: 19% (AC105204_2; initial mutation rate, 21%), 30% (AC105411, 37%), 41% (AC104013, 53%), 74% (AC093511, 76%), and 77% (AL442644, 92%; Table 3). There were no significant differences in mutation rates in the two sample sets (χ^2 test, P values: 0.20, 0.14, 0.08, 0.20, and 0.07, respectively). The seven coding repeats harbored mutation rates of 93% (*TGFβRII*, containing A10), 28% (*BLM*, A9), 27% (*CtIP*, A9), 54% (*MSH3*, A8), 27% (*MSH6*, A8), 26% (*IGFIIR*, G8), and 55% (*BAX*, G8), respectively (Table 3). None of the coding repeats showed length polymorphisms in germ line tissue. In 99% of mutant cases, mutant allele peak intensities were >30% of the corresponding wild-type signal.

Intronic repeats. Nine previously unpublished intronic C/G8 repeats, selected distant from exon-intron borders, showed somatic mutation frequencies between 11% and 22% (Table 4). In five of these repeats, germ line polymorphisms were detected (Table 4). Of nine intronic A8 and C8 repeats originally published by Zhang et al. (15), two loci failed to amplify, and in seven successfully analyzed repeats, somatic mutation frequencies of 10% to 47% were detected. Germ line polymorphism was detected in four of seven of the repeats (Table 5). In previous studies of nine intronic repeats, polymorphic samples were not excluded (15–17), and therefore, the mutation rates between these studies and our data are not directly comparable. To enable the comparison, we have determined the mutation rates of these seven repeats both by excluding and including polymorphic cases, and both of these mutation frequencies are presented in Table 5. Altogether, in 16 intronic repeats analyzed, 96% of all mutant peaks were >30% of the corresponding wild-type signal.

Table 2. Somatic mutation rates detected in 114 intergenic A/T and C/G repeats with 30 MSI CRCs

Repeat type	No. of repeats	Range of mutation rates (%)	Average mutation rate (%)	Cases with mutation rate ≥30%, no. of cases (%)
6 bp	29	0–7	0.55	0 (0/29)
A/T6	18	0–7	0.56	0 (0/18)
C/G6	11	0–3	0.55	0 (0/11)
7 bp	25	0–10	3	0 (0/25)
A/T7	14	0–10	3	0 (0/14)
C/G7	11	0–10	3	0 (0/11)
8 bp	22	0–37	13	9 (2/22)
A/T8	13	0–21	6	0 (0/13)
C/G8	9	10–37	24	22 (2/9)
9 bp	16	3–81	32	38 (6/16)
A/T9	10	3–53	22	30 (3/10)
C/G9	6	23–81	49	50 (3/6)
10 bp	22	0–92	50	78 (17/22)
A/T10	14	8–76	40	71 (10/14)
C/G10	8	0–92	67	88 (7/8)

Table 3. Somatic mutation rates in five intergenic repeats, and *TGFβRII, BLM, CtIP, TCF-4, MSH3, MSH6, IGFIIR*, and *BAX* analyzed with 100 MSI CRCs

	Repeat type	Coding/intergenic	In this study, no. of cases (%)	In previous studies (%)*
AC093511	A/T10	Intergenic	74 (67/90)	ND
AL442644	C/G10	Intergenic	77 (66/86)	ND
AC104013	A/T9	Intergenic	41 (40/98)	ND
AC105204_2	A/T8	Intergenic	19 (19/98)	ND
AC105411	C/G8	Intergenic	30 (29/96)	ND
TGFβRII	A10	Coding	93 (87/94)	60–94
BLM	A9	Coding	28 (27/96)	7–23
CtIP	A9	Coding	27 (24/90)	23–46
TCF-4	A9	Coding	59 (73/123)†	33–53
MSH3	A8	Coding	54 (48/89)	18–55
MSH6	C8	Coding	27 (25/92)	14–34
IGFIIR	G8	Coding	26 (23/90)	10–28
BAX	G8	Coding	55 (46/84)	33–63

Abbreviation: ND, not determined.
*See refs. (8, 14, 24).
† *TCF-4* sequencing was done in our previous study (18).

Discussion

To our knowledge, this is the first large-scale analysis of somatic background mutation rates in MSI cancer. In our recent pilot study, we screened 10 intergenic A/T9 repeats with MSI CRCs and detected high mutation rates in general, 70% being the highest rate observed. In that work, 60% (6 of 10) of A/T9 repeats showed a mutation frequency of ≥20%, and in 20% (2 of 10) of the repeats, a mutation frequency of ≥50% was detected (18). In this extended study of 114 intergenic repeats, the results of our pilot study were confirmed: High mutation rates up to 87% were detected in presumably neutral repeats. It has been previously known that mutation frequency in mononucleotide microsatellites is dependent on the length of the repeat, and C/G repeats are more prone to mutations than A/T repeats (14, 15, 23). Also in this study, a strong positive correlation (correlation = 0.81) between repeat length and mutation rate was detected, and long C/G repeats were significantly more unstable than A/T repeats. Therefore, our results were consistent with previous studies. It has been suggested that the surrounding sequence might affect the mutability of a microsatellite (15, 18). We did not detect a significant correlation between repeat instability and flanking sequence. In order to analyze the possible effect in more detail, a larger, and more homogenous repeat panel would be desired.

We wished to compare the mutation frequencies of intergenic repeats with the mutation frequencies of suggested MSI target genes by analyzing the coding repeats of the suggested MSI target genes. The seven coding repeats analyzed in this study (*TGFβRII, BLM, CtIP, MSH3, MSH6, IGFIIR*, and *BAX*), and one coding repeat (*TCF-4*) screened in our previous study (18), harbored similar or slightly higher mutation rates with 100 MSI CRCs described in previous studies (8, 14, 24). Seven previously published intronic A/T8 and C/G8 repeats (15–17) successfully amplified here showed mutation frequencies of 10% to 47% in our analysis after excluding all polymorphic cases. If this data needs to be compared with three previously published studies of intronic A8 and G8 repeats (15–17), mutation rates should be calculated by including polymorphic cases. Using this approach, the mutation rates here were significantly lower than those detected by Zhang et al. ($P \leq 0.001$, χ^2 test; ref. 15), in which cell lines and xenografts were used, and all the repeats were not analyzed in germ line tissue. However, compared with somatic mutation rates detected by Suzuki et al. and Duval et al. in primary tumors (16, 17), our sample set showed higher mutation rates ($P \leq 0.001$, χ^2 test). The differences in mutation rates can be explained by the different sample materials studied, and additionally, mutation detection instrumentation might be variable. Generally, the modern sequencing instrumentation we used in this study provides high resolution, higher than gel separation protocols used widely in the past. However, because the results of coding and selected intronic repeat sequencing in this study show consistency with previous studies, the high

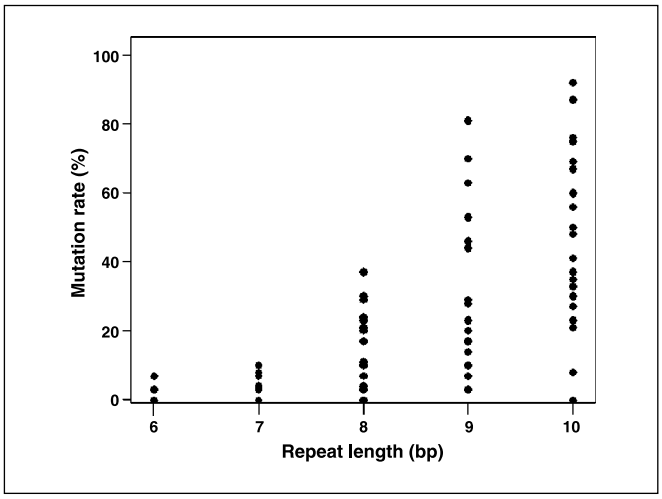

Figure 1. Correlation between mutation rate and repeat length in intergenic repeats with 30 MSI CRCs. Correlation between mutation rate and repeat length in all the repeats was 0.81 ($n = 114$, $P \leq 0.001$). *Black spheres,* data points. Correlation was calculated using Spearman correlation.

Table 4. Somatic mutation rates of nine unselected intronic C8 and G8 repeats with 30 MSI CRCs

Loci	Ensembl gene ID	Repeat	Distance to exon/intron border (bp)	Mutation rate, no. of cases (%)	Polymorphic samples, no. of cases
PPP1R9A	ENSG00000158528	C8	4,391	22 (6/27)	0/27
PRB4	ENSG00000121335	C8	1,531	17 (5/30)	0/30
Q8TEQ0_HUMAN	ENSG00000048471	C8	22,375	12 (3/26)	0/26
OKL38_HUMAN	ENSG00000140961	C8	144	14 (4/29)	0/29
PLEKHA5*	ENST00000299275	G8	9,922	11 (3/27)	3/30
STRN3*	ENSG00000196792	C8	9,922	14 (3/21)	9/30
DYNC1I1*	ENSG00000158560	G8	1,649	19 (4/21)	4/25
SPON1*	ENSG00000152268	C8	5,454	11 (2/18)	12/30
CD163L1*	ENSG00000177675	C8	7,323	22 (6/27)	4/30

*Repeats with length polymorphisms in germ line DNA.

intergenic repeat mutation rates detected here cannot be explained by methodologic factors.

There are only few studies of background mutation rates in MSI cancer, and from noncoding areas, only intronic repeats have been analyzed. To compare the mutation frequencies of unselected transcribed and nontranscribed noncoding repeats, we analyzed nine unpublished intronic C/G8 repeats. This comparison was of interest because published studies of unselected intronic repeats show low somatic mutation rates of ≤6% (8, 14). Also, our results suggest that overall, somatic mutation rates of unselected intronic repeats are low because 22% was the highest mutation rate detected in C/G8 repeats. In this study, intergenic C/G8 repeats harbored significantly more somatic mutations compared with the identical intronic repeats screened in this study, and in previous studies ($P \leq 0.01$, χ^2 test; ref. 14). For this comparison, we only selected studies of intronic repeats with mutations of verified somatic origin, and repeats of unselected nature to exclude possible selection bias. For these reasons, three studies (15–17) were excluded. The major difference between introns and intergenic regions is that the latter were not transcribed. Transcribed DNA is subject to transcription-coupled repair, which was primarily thought to remove UV-induced lesions. However, there is more recent evidence that transcription-coupled repair also removes oxidative DNA damage, and this could well be a source of mutations and other types of genomic instability (25). As far as repeats are concerned, introns contain the obligate polypyrimidine stretch, which is required for the correct splicing of pre-mRNA and which should be conserved. Cells mutated in these regions might be selected against. It is also possible that the frequency of homologous recombination between transcribed genes is higher than between nontranscribed regions. Thus, a mutation in a transcribed region may be repaired by gene conversion more readily than a mutation in a nontranscribed region. The latter point is unclear in normal genes, but class switch recombination in immunoglobulin loci is dependent on transcription (26). The putative mechanisms underlying the difference in correction of strand slippage between intergenic and intronic regions warrants further study.

Table 5. Somatic mutation rates of previously published intronic A8 and G8 repeats analyzed in this study (originally published by Zhang et al.)

Intronic repeats	Germ line	In this study, no. of cases (%)*	In this study, no. of cases (%)†	In Zhang et al., no. of cases (%)†,‡	In Suzuki et al. and in Duval et al., no. of cases (%)†,§
SMT4 (A8)	Monomorphic	34 (10/29)	34 (10/29)	29 (7/24)	12 (6/51)
SMT10 (A8)	Monomorphic	10 (3/30)	10 (3/30)	25 (6/24)	3 (1/52)
SMT28 (G8)	Monomorphic	23 (7/30)	23 (7/30)	29 (7/24)	6 (2/31)
SMT6 (G8)	Polymorphic	27 (6/22)	21 (6/28)	38 (9/24)	2 (1/52)
SMT15 (G8)	Polymorphic	27 (6/22)	20 (6/30)	46 (11/24)	7 (4/54)
SMT16 (G8)	Polymorphic	47 (7/15)	23 (7/30)	54 (13/24)	2 (1/58)
SMT29 (A8)	Polymorphic	12 (3/26)	10 (3/30)	38 (9/24)	0 (0/58)

*Polymorphic samples excluded from the study.
† Polymorphic samples included and scored for presence of somatic mutation.
‡ All the repeats were not examined in germ line tissue (15).
§ Taken together (16, 17).

In our study, intergenic repeats 8 to 10 bp were frequently mutated, and in MSI target gene studies, repeat sizes of 8 to 10 bp were also especially relevant. By 2002, 131 coding repeats length 6 to 10 bp had been investigated in MSI CRC, and the great majority of those repeats, 79% (99 of 131), were 8 to 10 bp (8). The generally accepted MSI target genes *TGFβRII* (A10) and *BAX* (G8) show mutation frequencies of 60% to 94%, and 33% to 66%, respectively, in previous studies in MSI CRCs (8). In this study, somatic mutation frequencies of 93% (*TGFβRII*), and 55% (*BAX*) were detected in 100 MSI CRCs. When candidate MSI target genes are evaluated, the type and length of the repeat in question is highly relevant, in addition to the observed mutation frequency. In addition to the high mutation rate, functional analyses could provide additional proof of the possible involvement in MSI tumorigenesis. In MSI target genes, frameshift mutations in the repeat cause impaired protein function, and hence, functional consequences might be detected using *in vitro* and *in vivo* models. For example *TGFβRII* is accepted as a real MSI target gene by functional modeling (13). However, *TGFβRII* defects have been subsequently proven to associate with Marfan syndrome according to genetic and *in vitro* analyses (27). Hence, mutations in one gene can be associated with distinct phenotypes. Another similar example is the strong association of *BRAF* mutations in developmental disorders—but not cancer—in the context of germ line mutation, despite the well-established role of *BRAF* as a somatically mutated cancer gene (28). Thus, when examining candidate MSI target genes, interpretation of functional data as well as *in vivo* phenotypes may be challenging, and even comprehensive approaches such as animal models may not always produce conclusive evidence.

Of the many criteria proposed in 1997, the most robust seems to be mutational involvement of the gene in MSS tumorigenesis. Although such involvement would show very strong evidence for selection, it is clear that MSI and MSS tumorigenesis follow somewhat different routes and true target genes are rejected if involvement in MSS tumorigenesis was considered a prerequisite. Thus, the identification of true MSI target genes is certainly possible if multiple lines of evidence support it, but many current candidate target genes must wait for formal recognition far into the future.

Acknowledgments

Received 11/27/2006; revised 3/7/2007; accepted 4/18/2007.

Grant support: Academy of Finland (Finnish Center of Excellence Program 2006-2011, grants 6302352 and 203610, to A. Karhu), the Finnish Cancer Society, the Sigrid Jusélius Foundation, the Association for International Cancer Research (grant 05-001), and the European Commission (LSHC-CT-2005-018754), and by grants from The Research and Science Foundation of Farmos, AstraZeneca, The Finnish Medical Foundation, Lilly Foundation, The University of Helsinki, The Finnish Cancer Society, and Biomedicum Helsinki Foundation (H. Sammalkorpi and P. Alhopuro).

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.

We thank Iina Vuoristo, Mikko Aho, Inga-Lill Svedberg, Mairi Kuris, Sini Marttinen, and Päivi Hannuksela for technical assistance.

References

1. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 1993;363:558–61.
2. Boland CR, Thibodeau SN, Hamilton SR, et al. 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 1998;58:5248–57.
3. Bronner CE, Baker SM, Morrison PT, et al. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature 1994;368:258–61.
4. Papadopoulos N, Nicolaides NC, Wei YF, et al. Mutation of a mutL homolog in hereditary colon cancer. Science 1994;263:1625–9.
5. Fishel R, Lescoe MK, Rao MR, et al. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 1993;75:1027–38.
6. Leach FS, Nicolaides NC, Papadopoulos N, et al. Mutations of a mutS homolog in hereditary non-polyposis colorectal cancer. Cell 1993;75:1215–25.
7. Duval A, Rolland S, Compoint A, et al. Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. Hum Mol Genet 2001; 10:513–8.
8. Woerner SM, Benner A, Sutter C, et al. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative real common target genes. Oncogene 2003; 22:2226–35.
9. Markowitz S, Wang J, Myeroff L, et al. Inactivation of the type II TGF-β receptor in colon cancer cells with microsatellite instability. Science 1995;268:1336–8.
10. Souza RF, Appel R, Yin J, et al. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. Nat Genet 1996;14:255–7.
11. Rampino N, Yamamoto H, Ionov Y, et al. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science 1997; 275:967–9.
12. Ionov Y, Yamamoto H, Krajewski S, Reed JC, Perucho M. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. Proc Natl Acad Sci U S A 2000;97:10872–7.
13. Wang J, Sun L, Myeroff L, et al. Demonstration that mutation of the type II transforming growth factor β receptor inactivates its tumor suppressor activity in replication error-positive colon carcinoma cells. J Biol Chem 1995;270:22044–9.
14. Vilkki S, Launonen V, Karhu A, Sistonen P, Vastrik I, Aaltonen LA. Screening for microsatellite instability target genes in colorectal cancers. J Med Genet 2002;39: 785–9.
15. Zhang L, Yu J, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B. Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. Cancer Res 2001;61:3801–5.
16. Suzuki K, Dai T, Suzuki I, Dai Y, Yamashita K, Perucho M. Low mutation incidence in polymorphic noncoding short mononucleotide repeats in gastrointestinal cancer of the microsatellite mutator phenotype pathway. Cancer Res 2002;6:1961–5.
17. Duval A, Reperant M, Hamelin R. Comparative analysis of mutation frequency of coding and non coding short mononucleotide repeats in mismatch repair deficient colorectal cancers. Oncogene 2002;21: 8062–6.
18. Hienonen T, Sammalkorpi H, Enholm S, et al. Mutations in two short noncoding mononucleotide repeats in most microsatellite-unstable colorectal cancers. Cancer Res 2005;65:4607–13.
19. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481–7.
20. Salovaara R, Loukola A, Kristo P, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol 2000;18:2193–200.
21. Wolfertstetter F, Frech K, Herrmann G, Werner T. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. Comput Appl Biosci 1996;12:71–80.
22. Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 1995;23:4878–84.
23. Boyer JC, Yamada NA, Roques CN, Hatch SB, Riess K, Farber RA. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet 2002;11:707–13.
24. Ikenoue T, Togo G, Nagai K, et al. Frameshift mutations at mononucleotide repeats in RAD50 recombinational DNA repair gene in colorectal cancers with microsatellite instability. Jpn J Cancer Res 2001;92:587–91.
25. Saxowsky TT, Doetsch PW. RNA polymerase encounters with DNA damage: transcription-coupled repair or transcriptional mutagenesis? Chem Rev 2006;106:474–88.
26. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol 2006;7:335–46.
27. Mizuguchi T, Collod-Beroud G, Akiyama T, et al. Heterozygous TGFBR2 mutations in Marfan syndrome. Nat Genet 2004;36:855–60.
28. Duesbery N, Vande Woude G. BRAF and MEK mutations make a late entrance. Sci STKE 2006;2006:pe15.

Background Mutation Frequency in Microsatellite-Unstable Colorectal Cancer

Heli Sammalkorpi, Pia Alhopuro, Rainer Lehtonen, et al.

Cancer Res 2007;67:5691-5698.