The Human Mutator Gene Homolog MSH2 and Its Association with Hereditary Nonpolyposis Colon Cancer

Richard Fishel, Mary Kay Lescoe, M. R. S. Rao, Neal G. Copeland, Nancy A. Jenkins, Judy Garber, Michael Kane, and Richard Kolodner

*Department of Microbiology and Molecular Genetics
Markey Center for Molecular Genetics
University of Vermont Medical School
Burlington, Vermont 05405

†Mammalian Genetics Laboratory
Advanced BioScience Laboratories Basic Research Program

‡Division of Cancer Epidemiology and Control
National Cancer Institute
Frederick Cancer Research and Development Center
Frederick, Maryland 21702

†‡Division of Cancer Epidemiology and Control
Dana–Farber Cancer Institute
Boston, Massachusetts 02115

§Division of Cellular and Molecular Biology
Dana–Farber Cancer Institute
and Department of Biological Chemistry
and Molecular Pharmacology
Harvard Medical School
Boston, Massachusetts 02115

Summary

We have identified a human homolog of the bacterial MutS and S. cerevisiae MSH proteins, called hMSH2. Expression of hMSH2 in E. coli causes a dominant mutator phenotype, suggesting that hMSH2, like other divergent MutS homologs, interferes with the normal bacterial mismatch repair pathway. hMSH2 maps to human chromosome 2p22-21 near a locus implicated in hereditary nonpolyposis colon cancer (HNPPC). A T to C transition mutation has been detected in the hMSH2 coding region of splice acceptor site in sporadic colon tumors and in affected individuals of two small HNPPC kindreds. These data and reports indicating that S. cerevisiae msh2 mutations cause an instability of dinucleotide repeats like those associated with HNPPC suggest that hMSH2 is the HNPPC gene.

Introduction

The faithful transmission of genetic information is paramount to the survival of a cell, an organism, and a species. Cells have evolved a number of mechanisms to ensure the high fidelity transmission of genetic material from one generation to the next since mutations can lead to genotypes that may be deleterious to the cell. The DNA lesions that lead to mutations are most frequently modified, missing, or mismatched nucleotides (Friedberg, 1985), and multiple enzymatic pathways have been described that specifically repair these lesions (Friedberg, 1990).

There are at least three ways in which mismatched nucleotides arise in DNA. First, physical damage to the DNA can give rise to mismatched bases (Friedberg, 1985). For example, the deamination of 5-methylcytosine creates a thymine and, therefore, a G-T mispair (Duncan and Millor, 1980). Second, incursion of nucleotides during DNA replication can yield mismatched base pairs and nucleotide insertions and deletions (Modrich, 1991). Finally, genetic recombination produces regions of heteroduplex DNA that may contain mismatched nucleotides when such heteroduplexes result from the pairing of two different parental DNA sequences (Holliday, 1964). Mismatched nucleotides produced by each of these mechanisms are known to be repaired by specific enzyme systems (Friedberg, 1990; Modrich, 1991).

The best-defined mismatch repair pathway is the Escherichia coli MutL system that promotes a long patch (approximately 2 kb) excision repair reaction that is dependent on the mutH, mutL, mutS, and mutU (uvrD) gene products (Modrich, 1989, 1991). The MutL holoenzyme appears to be the most active mismatch repair pathway in E. coli and is known both to increase the fidelity of DNA replication (Ryberg, 1978) and to act on recombination intermediates containing mispaired bases (Wagner and Mooreson, 1976; Fishel et al., 1986). This system has been reconstituted in vitro and requires the MutH, MutL, MutS, and UvrD (helicase II) proteins along with DNA polymerase III holoenzyme, DNA ligase, single-stranded DNA-binding protein, and one of the single-stranded DNA exonucleases (ExoI, ExoVII, or RecJ) (Modrich, 1989, 1991; Lahue et al., 1989; Cooper et al., 1993). MutS protein binds to the mismatched nucleotides in DNA (Su and Modrich, 1986). MutH protein interacts with GATC sites in DNA that are hemimethylated on the adenine and is responsible for incision on the unmethylated strand (Welsh et al., 1987). Specific incision of the unmethylated strand results in increased fidelity of replication because excision repair is targeted to the newly replicated unmethylated DNA strand. MutL facilitates the interaction between MutS bound to the mismatch and MutH bound to the hemimethylated Dam site, resulting in the activation of MutH (Grille et al., 1989). UvrD is the helicase that appears to act in conjunction with one of the single-stranded DNA-specific exonucleases to excise the unmethylated strand, leaving a gap that is repaired by the action of DNA polymerase III holoenzyme, single-stranded DNA-binding protein, and DNA ligase (Matson and George, 1987; Modrich, 1989, 1991; Lahue et al., 1989; Cooper et al., 1993). In addition, E. coli contains several short patch repair pathways (Fishel and Kolodner, 1989) including the very short patch (VSP) system (Lieb, 1987; Dzidic and Radman, 1989) and the MutY (MicA) system (Auer et al., 1988; Radicella et al., 1988), which act on specific single base mispairs.

While genetic studies on gene conversion in fungi were the first to suggest the existence of mismatch repair (Holliday, 1964), less is known about the enzymology of mismatch repair in eukaryotes than in prokaryotes. Genetic analysis suggests that Saccharomyces cerevisiae has a mismatch repair system similar to the bacterial MutLHS.
system (Williamson et al., 1985; Bishop et al., 1987; Reenan and Kolodner, 1992a, 1992b), even though S. cerevisiae lacks the type of DNA methylation that directs the strand exchanges of the MutHLS system (Proffitt et al., 1984). The S. cerevisiae pathway has a MutS homolog, MSH2 (Reenan and Kolodner, 1992a, 1992b), and, unlike the bacterial systems, two MutL homologs, PMS1 (Kramer et al., 1989) and MLH1 (Strand et al., 1993), that act in the same pathway. MSH2 has been shown to bind DNA containing mismatched nucleotides as predicted from its evolutionary relationship to the bacterial MutS proteins (Reenan and Kolodner, 1992a, 1992b; E. Alani, N.-W. Chi, and R. K., unpublished data). However, little is known about the other proteins that function in this pathway. S. cerevisiae has also been shown to contain other mismatch repair systems, including a mitochondrial mismatch repair system involving another MutS homolog, MSH1, which binds mismatched nucleotides (Reenan and Kolodner, 1992a, 1992b; N.-W. Chi and R. K., unpublished data). In addition, two other MutS homologs have been found in S. cerevisiae, MSH3 (New et al., 1993) and MSH4 (P. Ross-Macdonald and G. S. Roeder, unpublished data), although these two proteins are unlikely to play a major role in mismatch repair.

Biochemical studies have also provided evidence that eukaryotes have similar mismatch repair systems. Extracts of human (Holmes et al., 1990; Thomas et al., 1991), Drosophila (Holmes et al., 1990), and Xenopus (Varlet et al., 1990) cells can catalyze a mismatch repair reaction that resembles the bacterial MutLGS system. In addition, biochemical studies have found that human cells contain a nucleotide-specific mismatch repair system that appears to specifically recognize and repair G·T mismatches (Wiebauer and Jiricny, 1989, 1990). This latter system is similar to the E. coli MutY (MicA) system in that it involves a mismatch-specific DNA glycosylase (Au et al., 1989; Wiebauer and Jiricny, 1989, 1990).

In both bacteria and S. cerevisiae, mismatch repair plays additional roles in maintaining the genetic stability of DNA. The bacterial MutHLS system has been found to prevent genetic recombination between the divergent DNA sequences of related species such as E. coli and Salmonella typhimurium (termed homologous recombination) (Rayssiguier et al., 1989). A similar effect has been shown in S. cerevisiae, although the magnitude of the effect is not as large as that seen with bacteria (Bailis and Rothstein, 1990). Furthermore, genetic studies suggest that S. cerevisiae MSH2/PMS1-dependent mismatch repair controls the length of genetic exchanges (Alani et al., 1994). In this system, the length of heteroduplex regions formed during recombination appears to depend on the extent of mispairing between the two recombining sequences. Finally, S. cerevisiae msh2, pms1, and mih1 mutants have been found to exhibit increased rates of expansion and contraction of dinucleotide repeat sequences (Strand et al., 1993). This observation is of particular interest since different types of human tumors show an analogous instability of repeated DNA sequences (lonov et al., 1993; Thibodeau et al., 1993; Han et al., 1993; Risinger et al., 1993) and since hereditary nonpolyposis colon cancer (HNPCC) (also known as Lynch syndrome II [Bishop and Thomas, 1990; Lynch et al., 1991]) is linked to a locus causing a similar genetic instability (Aaltonen et al., 1993; Peltomaki et al., 1993).

We describe here the cloning and initial characterization of a human MutS homolog, hMSH2, and demonstrate that the gene maps to human chromosome 2p22-21, a region near the reported location of the HNPCC locus (Peltomaki et al., 1993). The similarity between S. cerevisiae msh2 mutants and HNPCC patients with regard to the instability of dinucleotide repeat sequences, the correspondence of the map locations of HNPCC and the hMSH2 gene, and the detection of a specific mutation in small HNPCC kindreds, as well as sporadic colorectal tumors, suggest that mutations in the hMSH2 gene are responsible for HNPCC.

Results

Cloning the Human MutS Homolog

To clone the hMSH2 gene, we utilized the degenerate polymerase chain reaction (PCR) approach that was successful in the identification of the S. cerevisiae MutS homologs MSH1 and MSH2 (Reenan and Kolodner, 1992a). Degenerate oligonucleotide primers were designed to target the amino acid sequences T6PMN and F(AT)Y(TM(F)) present in the most conserved regions of the known MutS homologs. One primer was used to target the invariant TAPNM sequence, whereas three primers were used individually to target FATH(FY), FVTH(FY), and FTTH(FY). The FATH(FY) sequence was of particular interest because it was most specific to the bacterial and S. cerevisiae homologs known to be involved in the major mismatch repair pathways in these organisms. These primers were used in a PCR containing human cDNA prepared from poly(A)+ RNA as template. PCR conditions were optimized to produce maximal yields of the expected 360 bp fragment, using both S. cerevisiae DNA and human cDNA as template. The resulting 360 bp fragment was then purified, and a number of independent clones were obtained and sequenced. One clone was identified using the FATH(FY) primer that contained an open reading frame encoding a predicted amino acid sequence that had 81% identity with the S. cerevisiae MSH2 protein and had an additional 10% conserved amino acid substitution. This cloned segment was used as a probe to screen a human cDNA library, and several apparent full-length cDNA clones were identified. A representative full-length clone had a 3111 bp insert that contained a 2727 bp open reading frame (Figure 1).

This open reading frame, when translated, was found potentially to encode a 909 amino acid-long protein that had 41% identity with the 966 amino acid S. cerevisiae MSH2 protein (Figure 2). The most conserved region, located between amino acids 573-764 of the human protein, was 85% identical to the corresponding region of the S. cerevisiae protein.

hMSH2 Is a Member of the MutS Mismatch Repair Protein Superfamily

Comparison of the hMSH2 amino acid sequence with the other known MutS homologs showed that it was most
The Human Mutator Gene MSH2

Figure 1. Sequence of the hMSH2 cDNA

The complete DNA sequence of the hMSH2 cDNA is given, starting with the first nucleotide present after the vector polylinker sequence and ending after the end of the poly(A) sequence found in the insert. Numbering is from the first nucleotide. Presented below the DNA sequence is the predicted hMSH2 amino acid sequence in single letter code, starting with the first in-frame methionine. The protein sequences targeted by the degenerate PCR primers are indicated by underlining.

highly related to S. cerevisiae MSH2 (Reenan and Kolodner, 1992a) and that these two amino acid sequences are more highly related than many pairs of the bacterial sequences (Figure 3). The evolutionary relationship among all of the known MutS-related proteins indicates that hMSH2 is a member of a group of MutS homologs that includes all of the bacterial homologs and S. cerevisiae MSH1 and MSH2. This group of proteins contains all of the members known to play a role in mismatch repair and in which mutations of the respective genes have been found to cause a strong mutator phenotype in these respective organisms.

Expression of hMSH2 Causes a Dominant Mutator Phenotype in E. coli

Expression of a divergent member of the bacterial MutS protein family in E. coli has been shown to cause a dominant mutator phenotype (Prudhomme et al., 1991). Con-
hMSH2 Is Located on Human Chromosome 2

Two methods of analyzing the National Institute of General Medical Science (NIGMS) mapping panel 2, containing chromosome-specific cell hybrid DNAs, have been used to determine the chromosomal location of the hMSH2 gene. The hMSH2 gene was localized to chromosome 2 by Southern blot analysis. A commercially obtained blot with mismatched nucleotides but lacks the ability to interact with other proteins in the repair pathway, thus interfering with normal mismatch repair. To gain an insight into whether hMSH2 plays a role in mismatch repair, the hMSH2 protein was expressed in E. coli under control of the lac promoter present in plasmid pMSH11, and the resulting cells were tested for an increased rate of accumulation of ampicillin plus rifampicin (rif') mutations. Mutations of E. coli that result in rif' have been mapped to the Ji subunit of RNA polymerase and have generally been found to have no effect on cell growth or viability (Nöno and Glass, 1982). Typical results of this mutator assay are shown in Figure 4. These results suggest that E. coli cells expressing hMSH2 accumulate rif' mutations at a higher rate than the isogenic wild-type strain containing the vector alone. Fluctuation analysis (Lea and Coulson, 1949) indicates that wild-type E. coli has a rif' mutation rate of 1.4 x 10^-6, that an isogenic strain of E. coli containing a mutS mutation has a rif' mutation rate of 4.8 x 10^-7, and that the wild-type E. coli expressing hMSH2 has a rif' mutation rate of 1.2 x 10^-6. Thus, expression of hMSH2 in E. coli results in an approximate 10-fold higher mutation rate. This observation is consistent with the idea that the hMSH2 protein can interact with mismatched nucleotides but lacks the ability to interact with other proteins in the E. coli MutHLS mismatch repair pathway; similar results were observed when Streptococcus pneumoniae HexpA was expressed in E. coli. (Prudhomme et al., 1991.)

The Mouse Homolog Maps to Mouse Chromosome 17 in a Region of Homology with Human Chromosome 22-21

The map location of the hMSH2 gene was further refined by recombination analysis using progeny derived from matings of (C57BL/6J x Mus spretus)F1 x C57BL/6J mice. This interspecific backcross mapping panel has been typed for over 1400 loci that are well distributed among all the autosomes as well as the X chromosome (Copeland and Jenkins, 1991). C57BL/6J and M. spretus DNAs were digested with several enzymes and analyzed by Southern blot hybridization for informative restriction fragment length polymorphisms using a human cDNA 360 bp hMSH2 probe to the most conserved region of MutS and its homologs. A 9.4 kb M. spretus HindIII restriction fragment length polymorphism (see Experimental Procedures) was used to follow the segregation of the MsH2 locus in backcross mice. The mapping results indicated that MsH2 is located in the distal region of mouse chromosome 17 linked to anti-phosphotyrosine immunoreactive kinase (Tik), mouse homolog 1 of Soa (Msos), and the lutropin-choriogonadotropin receptor (Lhcg). Although 159 mice were analyzed for every marker and are shown in the segregation analysis (Figure 6), up to 191 mice were typed for some pairs of markers. Each locus was analyzed in pairwise combinations for recombination frequencies using the additional data. The ratios of the total number of mice exhibiting recombinant chromosomes to the total number of mice analyzed for each pair of loci and the most

Figure 3. Phylogenetic Tree of MutS-Related Proteins

All protein sequences except S. cerevisiae MSH4 were retrieved from NCBI/GenBank release 78. The MSH4 sequence was provided by P. Ross-Macdonald and G. S. Roeder (Yale University, New Haven, Connecticut). The first 21 amino acids of S. cerevisiae MSH1 encoding the mitochondrial targeting sequence was removed from the MSH1 sequence prior to construction of the phylogenetic tree. We believe this sequence is unrelated to the enzymatic function of the MutS-like proteins and that its presence significantly alters the outcome of the alignment.
The frequency of spontaneous rif' E. coli was measured by spotting 30 μl of saturated bacterial cultures onto plates containing ampicillin (total cells) or ampicillin plus rifampicin (rif' mutants). All plates and cultures contained 0.5 mM isopropyl-β-D-thiogalactopyranoside to induce expression from the lac promoter. Three examples of plates exhibiting spontaneous rif' colonies arising from bacterial cultures grown from independent transformant colonies (or isolated colonies, in the case of the E. coli mutS mutants) are shown. Two independent cultures of E. coli AB1157 mutS show the magnitude of the mutator phenotype caused by a mutation inactivating the bacterial MutHLS mismatch repair system. The six other sectors of the top plate have independent cultures of wild-type E. coli AB1157 containing the Bluescript SK(+) vector without a cloned insert as a control for the frequency of spontaneous background mutations to rif'. The two bottom plates contain 16 independent cultures derived from independent transformants of E. coli AB1157 containing the pMSH1 plasmid that contains the hMSH2 gene downstream of the lac promoter.

We have recently found that the distal region of mouse chromosome 17 shares a region of homology with human chromosome 2p (summarized in Figure 6). In particular, Msos1 has been placed on human 2p22-21 and Lhcggr has been placed on 2p21. The mapping of Msh2 between Msos1 and Lhcggr suggests that the human homolog will map on the short arm of human chromosome 2 in the vicinity of band 21.

The HNPCC locus has been mapped near human chromosome 2p16-15 by linkage to microsatellite markers (Peltomaki et al., 1993). HNPCC showed a high degree of linkage to a microsatellite polymorphism marker D2S123. D2S123 has been mapped 5 cM distal to D2S5, which has been localized to 2p16-15 by in situ hybridization, linkage, and somatic cell hybrid analysis (Peltomaki et al., 1993). HNPCC showed no linkage to D2S136, which is 14 cM proximal to D2S123, and a small degree of linkage to D2S779, which is 9 cM distal to D2S779 (Peltomaki et al., 1993). Our analysis of the published data suggests that the linkage of HNPCC to D2S119 was possibly greater than implied (Peltomaki et al., 1993), indicating that
Figure 5. MSH2 Maps to Human Chromosome 2
(A) Southern analysis of an EcoRI digest of total genomic DNA isolated from human cells, Chinese hamster cells, and hybrid Chinese hamster cells containing human chromosome 2. (B) PCR amplification of a specific fragment of hMSH2 from NIGMS mapping panel 2 DNAs. C, Chinese hamster parental genomic DNA; M, mouse parental genomic DNA; H, human parental DNA; MW, pBR322 BstNI digest molecular weight markers, 383 bp and 121 bp; 1, 2, 22, X, Y, individual human–rodent cell hybrids containing chromosomes 1–22 and X and Y. Chromosomes 14, 16, 17, 20, and 21 are from mouse cell hybrids, and the remaining chromosomes are from Chinese hamster cell hybrids. The band marked Msh2 was amplified only when total human genomic DNA and genomic DNA from the human-hamster cell hybrid containing chromosome 2 were included in the PCR.

Figure 6. MshP Maps in the Distal Region of Mouse Chromosome 17
MshP was placed on mouse chromosome 17 by interspecific backcross analysis. The segregation patterns of MshP and flanking genes in 159 backcross animals that were typed for all loci are shown at the top of the figure. For individual pairs of loci, more than 159 animals were typed (see text). Each column represents the chromosome identified in the backcross progeny that was inherited from the (C57BL/6J × M. spretus)F1 parent. The closed boxes represent the presence of a C57BL/6J allele, and open boxes represent the presence of a M. spretus allele. The number of offspring inheriting each type of chromosome is listed at the bottom of each column. A partial chromosome 17 linkage map showing the location of MshP in relation to linked genes is shown at the bottom of the figure. Recombination distances between loci (in centimorgans) are shown to the left of the chromosome, and the positions of loci in human chromosomes, where known, are shown to the right. References for the human map positions of loci mapped in this study can be obtained from the Genome Data Base, a computerized data base of human linkage information maintained by The William H. Welch Medical Library of The Johns Hopkins University (Baltimore, Maryland).

HNPCC might map closer to 2p22-21 than the assigned 2p16-15 location suggested. We are trying to confirm that hMSH2 is linked to D2S123 by developing a microsatellite marker linked to hMSH2 so that linkage analysis with D2S123 can be performed.

Association of hMSH2 and HNPCC
HNPCC, also known as Lynch syndrome II and Muir–Torre syndrome (Bishop and Thomas, 1990), is one of the major types of hereditary colorectal cancer. HNPCC kindreds are defined as those in which three or more closely related family members in at least two successive generations have had histologically verified diagnosis of colorectal cancer, at least one of whom was diagnosed before 50 years of age. Other tumors that are associated with the syndrome occur elsewhere in the gastrointestinal tract, the female genital system, and the urinary tract.

The close association of the hMSH2 gene and the HNPCC locus led us to search for hMSH2 mutations in HNPCC families. Because large chromosome 2-linked HNPCC kindreds were not immediately available to us, we have taken the approach of analyzing both sporadic tumors and small HNPCC (Figure 7) and HNPCC-like kindreds for hMSH2 mutations using single-stranded conformational polymorphism analysis, heteroduplex analysis, and direct DNA sequencing (Orita et al., 1989; Tassabehji et al., 1992; Gromp, 1993). Initially, we examined seven exons covering the carboxy-terminal 48% of the MSH2 coding sequence. Because of the limited analysis used in these experiments, we have clearly not conducted an exhaustive search for MSH2 mutations.

A collection of 26 paired DNA samples extracted from random colorectal tumors and matched normal tissues were analyzed for the presence of dinucleotide (CA) repeat instability. Seven tumors were identified that showed dinucleotide repeat instability in at least one repeat locus. These seven tumor and paired normal DNA samples were
screened for MSH2 mutations. Two of the tumor DNAs were found to have a T to C transition mutation at the -6 position of the splice acceptor site of the intron, located at nucleotide position 2020 of the cDNA sequence, that was not present in the matched normal DNA sample (Figure 8). Similar T to C transition mutations within the poly-pyrimidine tract of mRNA splice acceptor sites, particularly short polypurine tracts, have been shown to affect mRNA splicing (Roscigno et al., 1993). The remaining sequence of this region, including the coding sequences between nucleotides 2020-2225 and the intron splice donor site was identical to the sequence of the cloned genomic region. The sequence downstream of nucleotide position 2225, was identical to the wild-type sequence. When seven additional paired DNA samples from the original collection of 26 in which the tumor DNA did not show dinucleotide (CA), repeat instability were examined, the normal DNA sequence was found in DNA from all seven tumors and all seven of the paired normal tissue samples. A larger study of colon tumors and other tumor types will be required to determine the frequency of this and other potential hMSH2 mutations in different tumors.

We have examined DNA from the blood of affected individuals from nine small HNPCC and HNPCC-like kindreds for mutations, including the transition mutation described above. None of these kindreds have been tested for linkage to chromosome 2. Mutations were found in two classic HNPCC kindreds. We have observed that all three affected individuals tested from family 1 (see Figure 7) and both affected individuals tested from family 2 (see Figure 7) are heterozygous for this T to C transition (Figure 8). Two individuals from family 1 who have not had cancer (ages 69 and 73) did not contain the mutation. We did not find a mutation in any of the other kindreds examined, but, as discussed above, we have not conducted an exhaustive search for mutations, leaving open the possibility that other MSH2 mutations exist in these kindreds. We are now collecting additional DNA samples from other family members to conduct a more extensive analysis of these kindreds. In addition, we are analyzing other kindreds, including large chromosome 2-linked kindreds, for the presence of this and other hMSH2 mutations to determine whether hMSH2 mutations and HNPCC cosegregate.

Discussion

We have described the cloning of a human cDNA encoding a protein that is highly homologous to the S. cerevisiae MSH2 protein and that appears to be a member of a group of related proteins that includes the bacterial MutS/HexA proteins and the S. cerevisiae MSH1 and MSH2 proteins. hMSH2 appears to be a member of a subgroup of these proteins, which have all been implicated as playing a role in the major mismatch repair pathway in these organisms. This is in contrast with the other two subgroups of MutS/HexA-related proteins, containing S. cerevisiae MSH3 and MSH4, respectively, for which there is little evidence that these proteins play a role in mismatch repair and in which mutations cause little if any mutator phenotype (New et al., 1993; P. Ross-Macdonald and G. S. Roeder, unpublished data). The observation that the expression of the hMSH2 protein in E. coli causes a dominant mutator phenotype similar to the heterologous expression of other divergent bacterial proteins provides support for the idea that hMSH2 functions in mismatch repair in humans. Because of the high degree of relatedness between hMSH2 and genes encoding proteins known to function in mismatch repair by recognizing mismatched bases, we postulate that hMSH2 protein plays a similar role in mismatch repair in humans.

In addition to playing a major role in preventing mutations during DNA replication, the MutHLS-type mismatch
Figure 8. DNA Sequence of the mRNA Splice Acceptor Site Region of the Intron at Nucleotide Position 2020 of the cDNA (a [top strand] and (d [bottom strand]) show the sequence of DNA from normal colon tissue; (b [top strand]) and (e [bottom strand]) show the sequence of DNA from a sporadic colorectal tumor showing dinucleotide (CA) repeat instability; (c [top strand]) and (f [bottom strand]) show the sequence of DNA from blood of individual III-1 of family 1 (Figure 7). The T to C transition is indicated by an asterisk and the intron/exon junction is indicated by an arrow. Analysis of the sequencing chromatograms of the blood DNA samples from affected individuals of the families shown in Figure 7 indicates that there is enough T signal in the C peak of the top strand and enough A signal in the G peak of the bottom strand to indicate that these individuals are heterozygous at this site. The T and A peaks are reduced relative to the C and G peaks owing to suppression of the T and A signals that occurs during sequencing through runs of Ts and As.

repair systems also appear to regulate different types of genetic stability (Rayssiguier et al., 1989; Strand et al., 1993). A striking example of this is that mutation of the msh2, mhl2, and pms1 genes all result in an increase in the rate of expansion and contraction of dinucleotide repeat sequences in S. cerevisiae of up to 700-fold (Strand et al., 1993). This phenotype is similar to the destabilization of microsatellite repeated sequences associated with different types of cancer (Ionov et al., 1993; Thibodeau et al., 1993; Han et al., 1993; Risinger et al., 1993), including HNPCC (Lynch et al., 1991; Aaltonen et al., 1993), and provided the rationale for considering the possibility that hMSH2 might be the HNPCC gene. The presently available data support the idea that the hMSH2 gene maps to human chromosome 2p22-21. This conclusion is based on experiments that locate hMSH2 on human chromosome 2, including experiments that map the mouse MSH2 homolog to a region of the mouse genome that is homologous to human chromosome 2p22-21 and on the association between a putative hMSH2 mRNA splice acceptor site mutation in both sporadic colon tumors and germline DNA from affected individuals of small HNPCC kindreds. Given the correspondence between the map location of hMSH2 and the map location of the HNPCC locus, our initial analysis of mutations in the hMSH2 gene, and the correspondence between the phenotypes of S. cerevisiae msh2 mutants and HNPCC, we suggest that mutations in the hMSH2 gene are responsible for HNPCC.

We have observed in both sporadic colon tumors and in germline DNA from affected individuals of HNPCC kindreds a T to C transition mutation at the −6 position of the splice acceptor site of the intron located at nucleotide position 2020 of the cDNA sequence. This is the splice acceptor site in the intron in front of the exon encoding the most conserved region of the protein. The observation of independent occurrences of the same mutation in two sporadic tumors and two unrelated families was unexpected. Our sample size is not large enough to prove definitively a link to HNPCC; however, the observation of this mutation in tumor tissue, but not normal tissue of the same individual, provides a strong indication that this base change represents a mutation rather than a polymorphism. A larger study of both sporadic tumors and HNPCC kindreds will be required to establish more firmly that the observed base change is not a polymorphism and to establish an association of this mutation and other hMSH2 mutations with both HNPCC and sporadic cancers. In this regard, we have recently examined two affected and two unaffected individuals from a chromosome 2-linked HNPCC family (Bishop and Thomas, 1990). Both affected individuals had the T to C transition mutation, whereas both unaffected individuals had the wild-type sequence (D. T. Bishop, N. Hall, J. Burn, M. K., and R. K., unpublished data).

The observation of a putative mRNA splice acceptor site mutation is intriguing. De la Chappelle, Vogelstein, and their collaborators have shown that HNPCC is apparently not associated with loss of heterozygosity (Peltomaki et al., 1993). This suggests that HNPCC mutations may be dominant. If the putative mRNA splice acceptor site mutation causes exon skipping or some other type of aberrant mRNA processing, then a truncated protein or a protein in which some internal sequences are deleted might result. The expression of such a protein might cause a dominant negative effect by interfering with the normal mismatch repair machinery, much like that observed when hMSH2 (Figure 4) and HexA are expressed in E. coli (Prudhomme...
et al., 1991). Testing this hypothesis will require proving that the mutation observed actually affects mRNA splicing, that a mutant protein is synthesized, and that it interferes with normal mismatch repair.

Several lines of evidence have suggested that cancers require multiple oncoprotein mutations to produce a metastatic tumor (Renan, 1993). For example, six independent mutations have been documented for the development of colon tumors (Fearon and Vogelstein, 1990). This number of mutations appears to be excessively high to be accounted for by the normal frequency of spontaneous mutations in cells. A role for hMSH2 in the development of cancer would suggest that a defect in the mismatch repair system may lead to an increased rate of accumulation of spontaneous mutations. This increased mutation rate and associated genome instability would then underlie the development of HNPCC and sporadic cancers. We further suggest that mutations in any of the other genes in an MSH2-dependent mismatch repair pathway could also cause a high frequency of mutagenesis that leads to the accumulation of such mutations. Furthermore, a common feature of tumors cells is the gross rearrangement of genomic DNA, including chromosome loss and rearrangement (Reichmann et al., 1981; Visscher et al., 1990). It is possible that a loss of hMSH2 function could increase the frequency of homeologous recombination between divergent DNA sequences, leading to chromosome rearrangements.

In summary, we have described a gene that is a candidate to be the HNPPC gene. Homologs of this gene encode a protein that functions in mismatch repair. Taken together, these statements make a number of predictions. First, the hMSH2 protein should specifically bind to DNA containing mispaired bases. Second, mutations in the hMSH2 gene should cosegregate with cancers in large HNPPC families. Third, mutations in the hMSH2 gene should be common in sporadic colon tumors and in other tumors that show dinucleotide repeat instability. It is also possible that mutations in the hMSH2 gene will be found in tumors showing the types of genetic instability associated with homeologous recombination events. We would like to point out that mutations in other hMSH2-dependent mismatch repair pathway genes might also cause the same effects as mutations in hMSH2. Finally, it is also possible that we may find mutations in the hMSH2 gene or related pathway genes to be associated with sporadic occurrences of inherited diseases, such as fragile X syndrome (Kuhl and Caskey, 1993), caused by the expansion of repeated sequences.

### Experimental Procedures

**Chemicals, Enzymes, Oligonucleotides, DNAs, Libraries, and Vectors**

Ultrapure Tris (acid and base), FDTA, MgCl2, MgSO4, NaCl, and analytical grade sodium citrate, KCl, potassium phosphate monobasic (K2HPO4), and sodium phosphate dibasic (Na2HPO4) were obtained from Amresco (Solon, Ohio). Ultrapure glycerol was obtained from Mallinckrodt, Incorporated (Paris, Kentucky). Decylribonucleotide triphosphates and ATP were purchased from Pharmacia LKB Biotechnology, Incorporated (Uppsala, Sweden). NIGMS mapping panel 2 DNAs were from Coriell Cell Repositories (Camden, New Jersey), and a Southern transfer of a BamHI digest of these DNAs used in preliminary experiments was from Oncor (Gaithersburg, Maryland). Gelatin was purchased from Sigma (St. Louis, Missouri). Restriction endonucleases, calf intestinal phosphatase, T4 polynucleotide kinase, and T4 DNA ligase were purchased from New England Biolabs, Incorporated (Beverly, Massachusetts). Taq polymerase was purchased from Perkin-Elmer Cetus (Norwalk, Connecticut). [y-32P]CTP and [y-32P]ATP were purchased from Amersham (Arlington Heights, Illinois). Oligonucleotides were synthesized on an Applied Biosystems (Foster City, California) 394 DNA synthesizer and were depurinated and purified by standard methods. PCR products of hMSH2 were inserted into BamHI-digested Bluescript SK(+) vector DNA (Stratagene, La Jolla, California) using standard methods. Isolation of the hMSH2 cDNA clone (pMSH11) was done by screening a HeLa S3 CDNA library constructed in the UniZap vector system (Stratagene, La Jolla, California). Plating and screening the library were performed according to the recommendations of the manufacturer.

**Cloning hMSH2 Using Degenerate PCR**

Degenerate oligonucleotides that would hybridize to DNA encoding two highly conserved regions of the known bacterial MutS/HexA and S. cerevisiae MSH proteins were designed along the lines described by Reenan and Kolodner (1990a). The following amino acid regions were selected: primer la, FATH(F/Y) (noncoding strand) 5'GGATCC (G/A)(A/T)A (G/A)TG (G/A/T/C)GT (G/A)C-3'; primer 1b, FTTH(F/Y) (noncoding strand) 5'ACCGGCATCC (G/A) (A/T)TG (G/A)(G/T/C)TG (G/A)C-3'; primer 1c, FVTH(F/Y) (noncoding strand) 5'ACCGGCATCC (G/A)(T/C)(A/T)AC (G/A)(G/T/C)AA-3'; primer 2, VFPYRM (coding strand) 5'GCGGGATCC AC(G/A/JT/C) GG(G/A/JT/C) AA (T/C) ATG-3'. The CCGGATCC sequence at the 5' end of each oligonucleotide is the BamHI restriction enzyme cleavage site added to facilitate cloning of the amplification product into the Bluescript SK(+) vector. PCR amplification of known MSH sequences from yeast genomic DNA was used to optimize the PCR conditions using primer 2 paired with either primer 1a, 1b, or 1c. PCR was performed in a 50 µl volume containing 10 mM Tris (pH 8.3), 50 mM KCl, 0.01% gelatin, 200 µM each of dGTP, dATP, dCTP, and dTTP, 1 U of Taq DNA polymerase, and 25 pmol of each degenerate primer. Multiple concentrations of MgCl2 were tested (1 mM, 3 mM, 5 mM, and 10 mM) for each primer pair as well as multiple concentrations of yeast genomic DNA or human cDNA (10 ng, 100 ng, and 1 µg). Human CDNA was prepared by standard methods from mRNA purified from MoMLL cells (Moore and Fishel, 1990) using the mRNA purification kit (Pharmacy, Uppsala, Sweden). The optimal method for amplification using these degenerate oligonucleotides on cDNA was found to be 26 cycles of denaturation for 1 min at 94°C, annealing for 2 min at 45°C, and polymerization for 5 min at 72°C.

After electrophoretic analysis of the products on a 2% agarose gel run in TAE buffer (45 mM Tris [pH 8.0], 5 mM sodium acetate, 2 mM EDTA), reactions that were deemed to contain products of the expected size (360 bp) were extracted with buffered phenol, precipitated in ethanol, and fractionated on a preparative 2% agarose–TAE gel containing 0.5 µg/ml ethidium bromide (Sigma, St. Louis, Missouri). The DNA band of interest was then isolated from the gel using NA45 paper essentially as described by the manufacturer (Schleicher & Schuell, Keene, New Hampshire) with the modification that the DNA was eluted from the NA45 paper by incubation at 70°C for 1 hr in 300 µl of 1 M NaCl, 50 mM arginine (free base). The elution solution was removed and extracted with buffered phenol, and the DNA was precipitated with ethanol. This isolated DNA fragment was digested with BamHI and isolated from a 2% agarose–TAE gel using NA45 paper, as described above, to remove the linker. The Bluescript SK(+) vector was digested with BamHI, treated with 20 U of calf intestinal phosphatase in a 50 µl reaction, and isolated from a 1% agarose gel using NA45 paper as described above. Fragment (20 ng) and Bluescript vector (200 ng) were added to a ligation reaction (100 µl) containing 50 mM Tris (pH 7.8), 8 mM MgCl2, 5 mM β-mercaptoethanol, 67 µM ATP, and 40 U of T4 DNA ligase and were incubated at 12°C for 16 hr, and then the DNA was transformed into E. coli XL1 blue (Stratagene, La Jolla, California) by the standard Mg++, Ca++ transformation procedure (Sambrook et al., 1989). Small-scale preparations of plasmid DNA (Sambrook et al., 1989) from individual transformants were analyzed for the presence of the appropriately sized insert (360 bp).
and 10 such clones generated with each primer pair were analyzed by double-stranded DNA sequencing. We found one MSH2 homolog among the 10 clones generated with the primer 1a plus primer 2a, and this plasmid was designated pMSH2.

The MSH2 homolog sequence contained in pMSH2 was used as a probe to screen a human cDNA library (UniZap HeLa S3 cDNA, Stratagene, La Jolla, California) according to the recommendations of the manufacturer. Urease nucleotide primers 15998 (5'-G3 UTTCTTTGTATGCT-3') and 15607 (5'-AGGACACCACTTGGTCG) were designed to hybridize to nucleotides inside the degenerate primer sequences on both ends of the MSH2 sequences present in pMSH2. A 278 bp fragment was amplified by PCR using these primers and was purified using NAAS paper as described above. A radiolabeled probe was made by performing PCR using 25 cycles of denaturation for 1 min at 94°C, of annealing for 2 min at 50°C, and of polymerization for 2 min at 72°C with a 50 µl reaction containing 1.5 mM MgSO4, 10 ng of the isolated 278 bp fragment, 200 µM each of dATP, dGTP, and dTTP, 25 pmol each of the two primers (15998 and 15607), and 100 µCi of [alpha-32P]dCTP (5000 Ci/mmol). Unincorporated nucleotides were removed by chromatography on anick column (Pharmacia, Uppsala, Sweden), the probe was denatured by boiling for 5 min, and 1 x 105 to 1 x 106 total disintegrations per minute were used to probe Hybond N filters (Amersham, Arlington Heights, Illinois) containing the UniZap HeLa S3 cDNA plate lifts (105 members). Two additional screens were carried out to isolate a homogenous lambda UniZap HeLa S3 cDNA phage population, and the insert was rescued using the R408 helper filamentous phage as described by the manufacturer (Stratagene, La Jolla, California). One positive clone containing a large 211 bp cDNA insert was used with a 2727 bp open reading frame homologous to MSH2 was characterized by DNA sequencing and designated pMSH11.

DNA Sequence Analysis

DNA sequencing of double-stranded plasmid DNAs was done with an Applied Biosystems (Foster City, California) 373A DNA sequencer using standard protocols and dye-labeled dideoxynucleoside triphosphates as terminators (Sanger et al., 1977; Smith et al., 1986). DNA sequence analysis was performed by using Sequencher 2.0.10 (Gene Codes Corporation, Ann Arbor, Michigan). Data base searches were performed at the National Center for Biotechnology Information using the basic local alignment search tool (BLAST) net service (Altschul et al., 1990). DNA and protein sequence homologies were determined using DNAStar MegAlign using the CLUSTAL method (Higgins et al., 1992). Multiple alignment parameters were a gap penalty of 10 and a gap length penalty of 10. Pairwise alignment parameters were a ktp value of 1.0, gap penalty of 3, window of 5, and a diagnol saved of 5. The phylogenetic tree was also constructed using DNAStar MegAlign (Ruthetailed and Nei, 1992).

Southern Hybridization

NIGMS mapping panel 2 DNAs were digested with EcoRI, and 10 µg of the resulting genomic DNA fragments was separated by electrophoresis through a 1% agarose gel run in TAE buffer. Southern transfer was performed according to Sambrook et al. (1989) onto Hybond N paper. Probe was prepared using the PCR method described above except that primers were used that amplify the full-length MSH2 fragment. We have found that this probe identifies EcoRI fragments containing the largest exons but does not identify any of the genomic EcoRI fragments containing MSH2 exons, presumably because of underrepresentation in the probe of some MSH2 sequences from the central portion of the insert (data not shown). The Southern blot data shown in Figure 5 were independently verified in experiments in which biotin was hybridized with different probes specific to subregions of MSH2 covering all of the MSH2 coding sequence (data not shown).

PCR Mapping

PCR was used to detect MSH2 sequences in the NIGMS mapping panel of DNAs using primers 16388 (5'-TTTCTTTGTATGCTC-3') and 16389 (5'-AAACCACTTGGTCGTTAG) that amplify a predicted 158 bp fragment of MSH2 contained in an intron located at nucleotide position 2020 of the cDNA sequence. PCRs (25 µl) contained 10 mM Tris buffer (pH 8.5), 50 mM KCl, 3 mM MgCl2, 0.01% gelatin, 50 µM each of dGTP, dATP, dTTP, and dCTP, 1.5 µl of Taq DNA polymerase, 5 pmol of each primer, and 0.5 µg of each DNA sample. PCR was performed for 20 cycles of denaturation for 30 s at 94°C, of annealing for 30 s at 55°C, and of polymerization for 1 min at 72°C, and 3 µl of each reaction was analyzed by electrophoresis through a 1.4% agarose gel run in TAE buffer.

Analysis of the MSH2 Gene for Mutations

The coding sequence from nucleotide positions 2120-2225 of the cDNA and the flanking regions of the genomic DNA containing the intron/exon junctions were amplified by PCR using primers 16324 (5'-GCCGATTACATCATGATG) and 16340 (5'-GGACACAGACATCATTCTATC). PCR was as described for PCR mapping except that the reactions contained 25 ng of DNA, 0.1 µCi of [alpha-32P]dCTP and were performed for 35 cycles. The resulting PCR products were then analyzed by single-stranded conformational polymorphism analysis and heteroduplex analysis using standard methods (Orita et al., 1989; Tassabehji et al., 1992; Gromp, 1993). Direct sequence analysis was performed by preparing unlabeled PCR products using the same method, purifying the PCR product by the GeneClean method (Rio 101, Incorporated, La Jolla, California), and then sequencing the products using primers 16324 and 16340 as described under DNA sequence analysis. DNA from normal/tumor pairs was analyzed for dinucleotide repeat instability using microsatellite markers D2S123 and D2S119 and standard methods.

Two types of DNA samples were analyzed. DNA samples prepared from the blood of both affected and unaffected individuals from small kindreds who appeared to have Lynch syndrome II and Muir-Torre syndrome were provided by F. Li and S. Verselis (Dana-Farber Cancer Institute, Boston). The identification of the tumor and normal DNAs was provided prior to analysis, but the diagnosis of individual tumor types was provided after the analysis was complete. A collection of matched DNAs extracted from a random assortment of colorectal tumors and adjacent normal tissue was provided by M. Barrett, J. M. Jessup, and G. Steele, Jr. (New England Deaconess Hospital, Boston). The identification of the tumor and normal DNAs was provided prior to analysis, but the diagnosis of individual tumor types was provided after the analysis was complete.

Interspecific Backcross Mouse Mapping

Interspecific backcross progeny were generated by mating C57BL/J [x M. spretus]F1 females and C57BL/6J males as described (Copeland and Jenkins, 1991). A total of 205 N2 mice were used to map the Msh2 locus (see text for details). DNA and Southern blot hybridization analysis was performed essentially as described (Jenkins et al., 1982). All blots were prepared with Zetabind nylon membranes (AMF Cuno, Lovato, Perret, France). The probe, a 360 bp human cDNA clone, was labeled with [alpha-32P]dCTP using a random-primed labeling kit (Stratagene, La Jolla, California), washing was done to a final stringency of 1.0 x SSC, 0.1% SDS at 65°C. A fragment of 12.5 kb was detected in HindIII-digested C57BL/6J DNA, and a fragment of 9.4 kb was detected in HindIII-digested M. spretus DNA. The presence or absence of the 9.4 kb M. spretus-specific HindIII fragment was followed in backcross mice. A description of the probe and the restriction fragment length polymorphisms for one of the loci linked to Msh2 and Mso1 has been reported (Webb et al., 1993). One locus not previously reported is TIK (Icely et al., 1991). The probe was a 1733 bp BamHI fragment of mouse cDNA that detected 14.0, 6.1, 3.7, and 1.5 kb fragments in Scal-digested C57BL/6J DNA and 7.3, 5.6, 2.9, 2.1, and 1.5 kb fragments in Scal-digested M. spretus DNA. The other locus not previously reported for this cross is Lho (McFarland et al., 1989). The probe was an 622 bp fragment of rat cDNA that detected major fragments of 13.5 and 8.3 kb in C57BL/6J DNA and of 8.3, 7.1, and 3.1 kb in M. spretus DNA following digestion with BglI. The M. spretus-specific restriction fragment length polymorphisms cosegregated and were followed in this analysis. Recombination distances were calculated as described (Green, 1981) using the computer program SPRETUS MADNESS. Gene order was determined by minimizing the number of recombination events required to explain the allele distribution patterns.
The Human Mutator Gene MSH2

str31, tpx33, supE44, λ was determined using a plate assay. The nuclease containing anadiscrpt (Stratagene, La Jolla, California) plasmid derivative pMSH11 was transformed into AB1157 according to the procedure of Fishel et al. (1986). Ampicillin-resistant transformants were selected and grown to saturation in LB buffer containing 100 μg/ml ampicillin and 0.5 mM isopropyl-β-D-thiogalactopyranoside. Dilutions of this culture were plated on LB plates containing 100 μg/ml ampicillin to determine the total number of viable cells containing the pMSH11 plasmid and on LB plates containing 100 μg/ml ampicillin plus 100 μg/ml rifampicin (Sigma, St. Louis, Missouri) to determine the total number of spontaneous rif' mutants in the culture. The rate of mutation was calculated according to Lea and Coulson (1949) using

\[ r = \frac{N}{M} \]

where \( r \) is the median number of rif' mutations in an odd number of independent cultures (usually 15) and \( M \) is the average number of rif' mutations per culture. \( M \) was solved by interpolation from the known \( r \) value and then used to calculate the mutation rate \( r \), where \( r = M/N \), where \( N \) is the final average number of viable cells.

Acknowledgments

M. R. S. Ft. is on sabbatical leave from the Indian Institute of Science, Bangalore, India. We thank Lorena Kallal, Tim Bishop, Eric Alani, Nai Wen Chi, Sarah Selig, Louis Kunkel, John Seidman, Eric Lander, Charles Stiles, Gerard Hudin, Fred Li, David Livingston, Greg Gilmar tin, and Myles Brown for technical advice, helpful discussions, and criticism. Debbie Gilbert and Brian Cho provided excellent technical assistance in the mouse mapping experiments. Sean Flaherty performed some initial PCR experiments, and Shawn Guerrette helped in the development of the mutator assay. PAUL MORRISON, Christine Earnabino, and Lori Wirth of the Dana-Farber Cancer Institute Molecular Biology Facility performed all DNA sequence analysis. We are particularly grateful to Fred Li, Chris Larkin, Sigatas Verselis, Michael Bennet, J. M. Jessup, and Glenn Steele, Jr., for their gifts of DNA samples from HNPPC kindreds and sporadic colorectal tumors and for help with pedigree analysis. This research was supported by grants CA56542 (to A. F.), HG00305 (now numbered GM50006) (to Ft. K.), Earabino, and Lori Wirth of the Dana-Farber Cancer Institute for help with pedigree analysis. This research was supported by grants and Cancer Center Core Grant CA06516 (to the Dana-Farber Cancer Institute) from the National Institutes of Health, by an American Cancer Society International cancer research fellowship from the International Union against Cancer (to M. R. S. Ft.), and by the National Cancer Institute (Department of Health and Human Services) under contract N01-C0-74101 with Advanced BioScience Laboratories (to N. G. C. and N. A. J.).

Received August 8, 1993; revised November 18, 1993.

References


