Recurrent germline mutation in *MSH2* arises frequently de novo

Darius C Desai, Janet C Lockman, Robert B Chadwick, Xin Gao, Antonio Percesepe, D Gareth R Evans, Michiko Miyaki, Siu Tsan Yuen, Paolo Radice, Eamonn R Maher, Fred A Wright, Albert de la Chapelle

Division of Surgical Oncology, The Ohio State University, 410 W 10th Avenue, N-924 Doan Hall, Columbus, OH 43210, USA
D C Desai

Human Cancer Genetics Program, Comprehensive Cancer Center, The Ohio State University, 420 W 12th Avenue, Suite 646, Columbus, OH 43210, USA
J C Lockman
R B Chadwick
X Gao
F A Wright
A de la Chapelle

Department of Internal Medicine, Universita di Modena, Via del Pozzo 71, 41100 Modena, Italy
A Percesepe

Department of Medical Genetics, St Mary's Hospital, Hathersage Road, Manchester M13 0JH, UK
D G R Evans

Department of Biochemistry, Tokyo Metropolitan Institute of Medical Sciences, 3-18-22 Honkomagome, Bunkyo-ku, Tokyo 113, Japan
M Miyaki

Department of Pathology, Queen Mary Hospital, The University of Hong Kong, Pokfulam, Hong Kong
S T Yuen

Department of Experimental Oncology, Istituto Nazionale Tumori, Via Venezian 1, 20133 Milano, Italy
P Radice

Section of Medical and Molecular Genetics, Department of Paediatrics and Child Health, University of Birmingham, The Medical School, Edgbaston, Birmingham B15 2TT, UK
E R Maher

Correspondence to:
Professor de la Chapelle, delachapelle-1@medctr.osu.edu

Revised version received 9 May 2000
Accepted for publication 12 May 2000

Abstract

***Introduction*—An intronic germline mutation in the *MSH2* gene, A→T at nt942+3, interferes with the exon 5 donor splicing mechanism leading to a mRNA lacking exon 5. This mutation causes typical hereditary non-polyposis colorectal cancer (HNPCC) and has been observed in numerous probands and families world wide. Recurrent mutations either arise repeatedly de novo or emanate from ancestral founding mutational events. The A→T mutation had previously been shown to be enriched in the population of Newfoundland where most families shared a founder mutation. In contrast, in England, haplotypes failed to suggest a founder effect. If the absence of a founder effect could be proven world wide, the frequent de novo occurrence of the mutation would constitute an unexplored predisposition.**

***Methods*—We studied 10 families from England, Italy, Hong Kong, and Japan with a battery of intragenic and flanking polymorphic single nucleotide and microsatellite markers.**

***Results*—Haplotype sharing was not apparent, even within the European and Asian kindreds. Our marker panel was sufficient to detect a major mutation arising within the past several thousand generations.**

***Discussion*—As a more ancient founder is implausible, we conclude that the A→T mutation at nt942+3 of *MSH2* occurs de novo with a relatively high frequency. We hypothesise that it arises as a consequence of misalignment at replication or recombination caused by a repeat of 26 adenines, of which the mutated A is the first. It is by far the most common recurrent de novo germline mutation yet to be detected in a human mismatch repair gene, accounting for 11% of all known pathogenic *MSH2* mutations.**

(*J Med Genet* 2000;**37**:646–652)

Keywords: *MSH2*; recurrent mutation; splice donor site of exon 5; founder mutation

The large "C" family from Newfoundland that allowed hereditary non-polyposis colorectal cancer (HNPCC) to be mapped to the short arm of chromosome 2[1] was subsequently shown to have an in frame cDNA deletion of *MSH2* comprising nucleotides 265-314.[2] Later, the genomic abnormality was identified as an exon 5 splice donor site mutation comprising an A→T change at the third nucleotide of the intron, abbreviated *MSH2* IVS5+3 A→T or A→T at nt942+3, and the same mutation was identified in two further North American HNPCC kindreds.[3] Subsequently, when four out of 33 kindreds in eastern England were found to have it,[4] a founding mutational event in an Anglo-Saxon ancestor appeared plausible. It was later shown that as many as 11 out of 41 HNPCC kindreds in Newfoundland had the same mutation, constituting 27% of all known HNPCC families.[5] In 1999, an in depth study of haplotypes in kindreds with the mutation from Newfoundland, England, and the USA was undertaken to determine its origin.[5] That study concluded that in Newfoundland, eight families out of 11 studied had identical haplotypes, suggesting a single founding event. As Newfoundland began to be settled as recently as 1610, this finding was consistent with the observed conserved haplotype extending over a large region of 10 cM. In contrast, no clear evidence of haplotype sharing was observed among three US and five English families, prompting the suggestion that the mutation was unlikely to be the result of a world wide ancestral event.[5]

The rationale for undertaking the present investigation was that the same mutation began to be reported frequently in Europe and Asia.[3 6–14] By early 2000, a total of 114 different germline mutations of *MSH2* had been listed in the HNPCC database (www.nfdht.nl), and the A→T at nt942+3 was frequently observed. Indeed, considering both the database and the publications cited above, it is by far the most commonly reported *MSH2* germline mutation. It currently accounts for 11% of all *MSH2* germline mutations reported in the HNPCC database. If this were an ancestral mutation that occurred on several continents including Asia, its origin would have to be quite ancient. It follows that, as a result, the putative conserved and shared haplotype would only comprise a very small chromosomal region. Well known examples of conserved ancestral haplotypes occurring world wide include the region around the β globin gene on chromosomes with the Glu→Val missense mutation in codon 6 leading to sickle cell disease,[15] and the region of the *CFTR* gene on chromosomes with the ΔF508 mutation responsible for cystic fibrosis.[16] In both cases the mutation occurs world wide and accounts for a major proportion of the respective disease chromosomes.

Table 1 Characteristics of subjects studied

Family	*Identification*	*Relationship*	*A→T at nt942+3*
Italy (Milan 1)	1560	Son	Yes
	1561	Daughter	No
	1581	Son	Yes
	1587	Mother	No
	1729	Son	No
	2883	Daughter	No
Italy (Milan 2)	2417	Sister	No
	2517	Cousin	No
	2719	Cousin	Yes
	2412	Brother	Yes
	3045	Daughter of 2412	Yes
Italy (Modena)	1	Father	Yes
	2	Child	No
	3	Child	No
	4	Child	No
	5	Child	No
England family 1	1	Sib	Yes
	2	Sib	No
England family 2	1	Sib	Yes
	2	Sib	Yes
	3	Sib	Yes
England family 3	1	Mother	Yes
	2	Daughter	Yes
	3	Daughter	Yes
	4	Daughter	No
	5	Son	Yes
England family 4	1	Mother	Yes
	2	Daughter	No
	3	Daughter	Yes
	4	Daughter	No
	5	Distant relative	Yes
	6	Distant relative	No
Hong Kong 1	Reg 41	Sib	Yes
	Reg 49	Sib	No
	Reg 99	Sib	No
Hong Kong 2	ST99	Daughter	Yes
	ST114	Sib	No
	ST115	Father	Yes
Japan	2496	Grandmother	No
	1826	Son	Yes
	2497	Grandson	No
	2498	Grandson	Yes

Allele sharing at flanking marker loci can be used to estimate the number of generations since founding, provided that the genetic or physical distance between the mutation and marker loci is known.[17 18] Indeed, using this strategy, the age of the ΔF508 mutation in *CFTR* was tentatively determined as some 54 000 years or over 2000 generations.[16] By analogy, if the A→T at nt942+3 mutation of *MSH2* were an ancient founding mutation that occurs world wide, the shared haplotype would be predicted to be quite small.

If it could be convincingly shown that the A→T at nt942+3 mutation was not inherited from one or a few common ancestors, it would constitute an interesting repeated de novo occurrence in need of an explanation. The present study was designed to settle the question by revisiting previous marker and haplotype data by studying the inheritance of several polymorphic markers inside and immediately adjacent to the *MSH2* gene in patients and families with the mutation.

Materials and methods

PATIENTS

DNA from the blood of 42 members of 10 families with the A→T at nt942+3 mutation was obtained through clinics in Europe and Asia (three Italian, four English, two Hong Kong, and one Japanese). The Italian families from Milan were ascertained through the Hereditary Colorectal Tumor Registry of the National Cancer Institute of Milan, Italy. The characteristics of the kindreds are shown in table 1. All families satisfied the Amsterdam criteria for HNPCC families.

MARKER IDENTIFICATION AND MAPPING

Intragenic single nucleotide polymorphisms (SNPs) were identified using the web site of the International Collaborative Group on Hereditary Non-Polyposis Colon Cancer at www.nfdht.nl/database/msh2-poly.htm. Three SNPs were identified as likely to be informative, each with allele frequencies at the lesser allele of greater than 20% (c→g at 211+9, t→a at 1511-9, g→a at 1661+12). In addition, six microsatellites within or near *MSH2* were identified. Of these, the three tetranucleotide repeat microsatellites are intragenic and mapped relative to the exons, while the remaining three dinucleotide repeat microsatellites (CA) have not been precisely mapped, but have been sequenced on a BAC containing the *MSH2* gene. These three CA repeat sequences are within 70 kb of *MSH2*. Parametric methods for assessing linkage disequilibrium depend on genomic sequence position, but many introns of *MSH2* have not yet been fully sequenced. We used cross_match version 0.990315 (P Green, unpublished data, www.phrap.org) to compare *MSH2* mRNA and DNA (accession number AC009600), and to obtain relative genomic positions for the markers where possible. These markers were typed in all available family members. Another set of 16 SNPs were identified for study, but based on published reports are thought to be relatively uninformative. Nonetheless, a fortunate pairing of a disease allele with a low frequency marker allele may provide powerful evidence for linkage disequilibrium. These SNPs were thus typed for two subjects in each kindred, in either parent child pairs (seven kindreds) or sib pairs (three kindreds) to determine if further typing was warranted. All 16 markers were homozygous for the major allele in allele typed subjects and thus were not informative for this study.

MOLECULAR GENETIC ANALYSIS

SNP sequencing

Single nucleotide polymorphisms were genotyped by direct sequencing of genomic PCR products. PCR reactions were done in 25 μl volumes with 100 nmol/l of each of the respective PCR primers (table 2), 25 ng of genomic DNA, 100 μmol/l of each dNTP, 1.0 U Ampli*taq* Gold DNA polymerase (Perkin-Elmer, Norwalk, CT), 10 mmol/l pH 8.3 Tris-HCl, 50 mmol/l KCl, and 2 mmol/l $MgCl_2$. PCR fragments were purified using the Exonuclease I/Shrimp Alkaline Phosphatase PCR Product Presequencing Kit (USB-Amersham Life Science). After purification according to the manufacturer's protocol, 2 μl of the PCR products were sequenced using the BigDye Terminator Ampli*Taq* FS Cycle Sequencing Kit (PE Biosystems, Foster City, CA). The method for sequencing has been previously described.[19]

Microsatellite analysis

The primers used are listed in table 2. Amplifications were done in 15 μl PCR reaction volumes. Concentrations of the following

reagents were used: 1 µl of each 8 µmol/l primer (the 5' primer is fluorescently labelled), 10 ng of genomic DNA, and 8 µl of HotStar*Taq* Master Mix (Qiagen). The following thermal cycling profile was used: one cycle of 95°C for 12 minutes, followed by 35 cycles of 95°C for 10 seconds, 55°C for 15 seconds, and 72°C for 30 seconds, followed by one final extension of 72°C for 30 minutes, followed by a soak at 4°C. Respective PCR reactions for each marker were pooled together and loaded onto the PE377 automated sequencer. Allele sizing and calling was done using Genotyper software (PE Biosystems).

HAPLOTYPE ANALYSIS AND CONTROL HAPLOTYPES

Haplotype analysis was performed in each kindred using the GENEHUNTER linkage program,[20] augmented by inspection and direct Bayesian calculation. In each kindred the haplotype associated with disease was identified and any additional unambiguous haplotypes were used to form a control sample. The use of control haplotypes within the same kindreds can be important to reduce biases owing to population stratification,[21] and the use of external control genotypes would require the assumption of linkage equilibrium in order to construct haplotype frequencies.

STATISTICAL ANALYSIS

Markers were examined individually for evidence of allelic association with disease. In most instances, a unique allele could be identified as in phase with the disease mutation. Otherwise, both alleles in a genotype were included as disease alleles, which does not affect the type I error in the analysis. Allelic association with disease was assessed by Fisher's exact test. Additional multipoint linkage disequilibrium analyses were performed, using the SNP data alone and in combination with the microsatellite data. One of these analyses was a non-parametric test in which haplotypes of varying widths were examined for association with disease. For each haplotype of a fixed width, a chi-square test was performed on the contingency table of haplotype by disease status,[22] and the maximum value of the statistic over all haplotype widths

Table 2 Single nucleotide polymorphic markers used

SNP	*Primers*
+9 of 3' end of exon 1	Fw GACCGGGGCGACTTCTATAC
c→g at 211+9	Rev CACTCCGTGATCACAAGTTCAG
−9 of 5' end of exon 10	Fw AAACTAACATTCATAAGGGAGTTAAGG
t→a at 1511-9	Rev CCAAACTGTGCACTGGAATC
+12 of 3' end of exon 10	Fw TGTAAGGAAGAAAAAGTCCTTCG
g→a at 1661+12	Rev GAAAGCTTGACTCTTACCTGATGAC
+3 of 3' end of exon 5	Fw TGTAAAACGACGGCCAGTGTGGTATAGAAATCTTCGATTTTT
A→T at nt942+3	Rev CAGGAAACAGCTATGACCAATCAACATTTTTAACCCTTTT
Exon 2 codon 113	Fw TGTAAAACGACGGCCAGTGTCCAGCTAATACAGTGCTTG
AAG→AAA	Rev CAGGAAACAGCTATGACCCACATTTTTATTTTTCTACTCTTAA
Exon 2 codon 110	Fw TGTAAAACGACGGCCAGTGTCCAGCTAATACAGTGCTTG
A→G at 329	Rev CAGGAAACAGCTATGACCCACATTTTTATTTTTCTACTCTTAA
Exon 2 codon 96	Fw TGTAAAACGACGGCCAGTGTCCAGCTAATACAGTGCTTG
G→A at 287	Rev CAGGAAACAGCTATGACCCACATTTTTATTTTTCTACTCTTAA
Exon 3 codon 133	FwTGTAAAACGACGGCCAGTAAAGTATGTTCAAGAGTTTGTTAAATTTTT
C→T at 399	Rev CAGGAAACAGCTATGACCCTTTCCTAGGCCTGGAATCTC
Exon 3 codon 153	FwTGTAAAACGACGGCCAGTAAAGTATGTTCAAGAGTTTGTTAAATTTTT
C→T at 459	Rev CAGGAAACAGCTATGACCCTTTCCTAGGCCTGGAATCTC
Exon 6 codon 322	Fw TGTAAAACGACGGCCAGTTTTTCACTAATGAGCTTGCCATTCTT
G→A at 965	Rev CAGGAAACAGCTATGACCGGTAACTGCAGGTTACATAAAACTAACGA
Exon 6 codon 328	Fw TGTAAAACGACGGCCAGTTTTTCACTAATGAGCTTGCCATTCTT
GCC→GCT	Rev CAGGAAACAGCTATGACCGGTAACTGCAGGTTACATAAAACTAACGA
−10 of 5' end of exon 7	Fw TGTAAAACGACGGCCAGTCGACTTAGTTGAGACTTACGTGC
t→c at 1077-10	Rev CAGGAAACAGCTATGACCTTTATGAGGACAGCACATTGC
+49 of 3' end of exon 7	Fw TGTAAAACGACGGCCAGTCGACTTAGTTGAGACTTACGTGC
c→a at 1276+49	Rev CAGGAAACAGCTATGACCTTTATGAGGACAGCACATTGC
Exon 11 codon 556	Fw TGTAAAACGACGGCCAGTCATTGCTTCTAGTACACATTT
T→C at 1666	Rev CAGGAAACAGCTATGACCCAGGTGACATTCAGAACATTA
Exon 11 codon 579	Fw TGTAAAACGACGGCCAGTCATTGCTTCTAGTACACATTT
A→G at 1737	Rev CAGGAAACAGCTATGACCCAGGTGACATTCAGAACATTA
Exon 11 codon 585	Fw TGTAAAACGACGGCCAGTCATTGCTTCTAGTACACATTT
TCT→TCC	Rev CAGGAAACAGCTATGACCCAGGTGACATTCAGAACATTA
Exon 12 codon 596	Fw TGTAAAACGACGGCCAGTTCAGTATTCCTGTGTACATTT
A→G at 1787	Rev CAGGAAACAGCTATGACCTTACCCCCACAAAGCCCAA
−6 of 5' end of exon 13	Fw TGTAAAACGACGGCCAGTGTAGCAGAAAGAAGTTTAAAACTTGC
t→c at 2006-6	Rev CAGGAAACAGCTATGACCGGACAGAGACATACATTTCTATCTTC
Exon 13 codon 713	Fw TGTAAAACGACGGCCAGTGTAGCAGAAAGAAGTTTAAAACTTGC
G→C at 2139	Rev CAGGAAACAGCTATGACCGGACAGAGACATACATTTCTATCTTC
Exon 13 codon 718	Fw TGTAAAACGACGGCCAGTGTAGCAGAAAGAAGTTTAAAACTTGC
A→G at 2154	Rev CAGGAAACAGCTATGACCGGACAGAGACATACATTTCTATCTTC
Microsatellites	Primers
TTTA repeat	Fw FAM-GAGGATGGCCACAAATTAGC*
	Rev ACCAGGGAGTCAGAGTTTGC
TTTG repeat	Fw NED-TGGTGTAGGCAGCCATGTATC*
	Rev CCTCTGCCTAGGAAAACCAG
TAAA repeat	Fw HEX-TGAGTATTGCTCTCTTGCTATCTTG*
	Rev AGAGCCGTAATCACTCAATGTG
CA repeat 1	Fw FAM-TCACCCCAGCCAGACTCTAAG*
	Rev GTTTCTTCAGATTTTTATTGAGAACCTACCAGG
CA repeat 2	Fw FAM-TGTTGGACTCCGCAGGATTG*
	Rev GTTTCTTTAGGTGTATGTAGTAGAGGGCAAGC
CA repeat 3	Fw HEX-GTCTCTCTCTGTGTGTCTTTCTGCC*
	Rev GTTTCTTGCCAACTGGTTCCATTTGACC

*FAM, NED, and HEX are fluorescent labels tagged onto the 5' end of the custom primers.

Table 3 Distribution of alleles and tests for evidence of linkage disequilibrium at individual marker loci

	CA repeat(2) (p=0.40)				*CA repeat(3) (p=0.72)*				
Markers	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	
Disease	4	0	3	7	10	0	4	0	
Control	3	2	2	14	11	2	7	1	
	CA repeat(1) (p=0.55)						*SNP (213+9) (p=0.71)*		
Markers	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	*Allele 5*	*Allele 6*	*Allele 1*	*Allele 2*	
Disease	1	2	4	4	3	0	11	3	
Control	0	4	10	4	2	1	15	6	
	TTTA repeat (p=0.28)				*TTTG repeat (p=0.49)*				
Markers	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	
Disease	2	7	1	4	3	5	6	0	
Control	5	11	4	1	1	9	9	2	
	SNP (1512−9) (p=1)		*SNP (1662+12) (p=0.73)*		*TAAA repeat (p=0.67)*				
Markers	*Allele 1*	*Allele 2*	*Allele 1*	*Allele 2*	*Allele 1*	*Allele 2*	*Allele 3*	*Allele 4*	*Allele 5*
Disease	12	2	9	5	1	6	2	3	2
Control	19	2	11	10	0	7	2	6	6

p values calculated by Fisher's exact test.

was recorded. Overall significance was assessed by generating 1000 random permutations of the status (disease *v* control) of the haplotypes to create an empirical null distribution for the overall statistic. In addition, the program DISMULT[23] was used to assess the evidence for linkage disequilibrium. Because three of the markers are not precisely mapped and because SNPs are thought to have a much lower mutation rate than microsatellites, five separate multipoint analyses were performed: (1) all markers used**,** unmapped markers placed 5' of *MSH2*, (2) all markers used, unmapped markers placed 3' of *MSH2*, (3) using the six mapped markers only, (4) using the unmapped markers only, and (5) using SNPs only.

Results

In the 10 families, as we expected, there was at most one recombination in the markers in any of the kindreds, based on a 5' placement of the unmapped markers to *MSH2*. Table 3 presents the disease allelic association evidence at each marker locus. None of the markers shows even marginally significant association with disease. Allelic association with disease can be detected more powerfully based on identification of ancestral founder haplotypes, rather than relying on individual marker genotypes.[24] In seven of the kindreds the identification of the disease haplotype was unambiguous, because at least two generations of probands were typed or a sufficient number of affected sibs were typed. In two of the remaining kindreds (England 2 and Hong Kong 1), the phase of one of the mapped markers could not be known with certainty, and the disease haplotype was chosen according to an approximation to the maximum posterior probability.[20] In another kindred (England 1), the phases of two of the mapped markers were uncertain, and no reconstruction was attempted. However, this kindred was informative for haplotype analysis using SNPs alone. Analyses performed without these last three kindreds do not alter the conclusions of our study.

Table 4 describes the haplotypes observed in the kindreds. Ambiguous haplotypes for the unmapped markers are left blank. None of the five multipoint linkage disequilibrium analyses described in Methods was significant on these data. Using our non-parametric analysis, empirical p values ranged from 0.21 to 0.96. Both the DISMULT program and the p value approximation suggested by Terwilliger[23] yielded p values all in excess of 0.9. The lack of association with disease is apparent from tables 3 and 4. Using simple formulae describing decay of linkage disequilibrium,[24] we can explore the plausibility of a world wide founding event. From tables 3 and 4, we can rule out founding events within the past few thousand generations as responsible for the majority of A→T at nt942+3 mutations. For example, three of the markers are within about 20 kb of the mutation site, and historical recombinations over such an interval would occur at a rate of only about 10% in 500 generations, assuming the average correspondence 1 cM = 1 Mb. Thus, such a major haplotype would be apparent in at least the European kindreds. A major ancestral haplotype arising 2000 generations (approximately 50 000 years) ago would still be apparent in 45% of chromosomes at a distance of up to 40 kb. A single even more ancient founding event might fail to be recognised as such, because of recombinations between the mutation and the nearest markers. However, we deem it likely that such an ancient single ancestral mutation would still give rise to more recent founding haplotypes that would be apparent in subpopulations of more recent lineage. Again, we see no such evidence from our data, but we cannot entirely rule out an ancient founding event giving rise to multiple derived haplotypes.

The mutation rates of SNPs are thought to be much lower than of microsatellites, in the order of 10^{-9} per meiosis,[25] and we analysed the data using the SNP haplotypes alone. The second and third SNPs are separated by only 160 base pairs and are thus sensitive for detecting linkage disequilibrium, because recombinations between the two SNPs will be exceedingly rare. Thus these SNPs may be treated as a

Table 4 Haplotypes

Family	Disease haplotypes		Control haplotypes	
	Mapped markers	*Unmapped markers*	*Mapped markers*	*Unmapped markers*
Italy (Milan 1)	142122	116	113115	416
			112114	415
			122122	414
Italy (Milan 2)	141112	416	132112	
			133114	415
			122112	415
Italy (Modena)	112215	414	133123	436
			112115	134
England 2* England 3	123113		122112	
	132112	311	111225	133
			122112	223
			132112	414
England 4	213224	135	123113	234
			223124	435
			122112	414
Hong Kong 1*	143113		123124	
			223124	
Hong Kong 2	223125	434	223124	314
Japan	121112	413	224124	417
			223125	434
			243124	444

Disease and control haplotypes were identified as described in Methods. The mapped markers are SNP 1 (+9 3' exon 1, C to G), microsatellite 6890, microsatellite 36974, SNP 2 (−9 5' exon 10, A to T), SNP 3 (+12 3' exon 10, G to A), and microsatellite 62153. The unmapped markers are microsatellites CA(2), CA(3), and CA(1), in the order they appear on the *MSH2* BAC. CA(3) is approximately 20 kb from CA(2), and the set of unmapped markers is thought to be approximately 70 kb upstream from the 5' end of *MSH2*. SNP alleles are coded 1 or 2, with 2 being the less frequent allele. The A→T at nt942+3 mutation lies between the second and third of the mapped markers, at +3 3' of exon 5. Haplotype reconstruction was not reliable for family England 1, but this family is included in tables 3 and 5.
*Reconstructed haplotypes.

Table 5 SNP haplotypes

SNP haplotypes	*Disease haplotype source*	*No disease*	*No control*
111	1 Italian, 3 English, 1 Hong Kong, 1 Japanese	6	13
112	1 Italian	1	2
121	1 Italian	1	0
122		0	1
212	1 Hong Kong	1	6
222	1 English	1	0

p=0.38, Fisher's exact test.

single, more informative locus. Table 5 shows the SNP haplotypes alone and the corresponding families. Again there is no overall evidence of association with disease.

Discussion

The HNPCC Mutation Database (www.nfdht.nl) currently lists a total of 281 different germline mutations believed to be responsible for HNPCC. The listing is not complete but does reflect our present understanding of the mutation spectrum. Nevertheless, there are several good reasons to predict that the spectrum will change. First, the methods used to detect mutations have different sensitivities for different types of mutations. For example, hard to detect deletions are not uncommon.[26 27] Also, the absence of gene transcript owing to genetic changes not detectable by current routine methods may account for a sizeable proportion of all mutations.[27] Additionally, HNPCC may result from mutations in genes that have not yet been identified. Finally, as population based studies are becoming feasible, mismatch repair gene mutations are being detected in affected subjects that do not belong to large families or are entirely "sporadic".[28] Mutations ascertained in this way may have a different spectrum.

Despite the above shortcomings, the mutations listed in the database and relevant publications can be relied on to some extent to assess the existence of recurrent mutations. Among a total of 114 different germline mutations reported in *MSH2*, the great majority have been seen only once (in one patient or kindred). Some 16 have been seen in two to four patients or kindreds, often from the same country. Most likely, all of these represent relatively close genealogical kinships so that the mutation is derived from a not too distant common ancestor. The same may or may not be true of the AAT deletion at codon 596 in exon 12 that has been reported on at least seven different occasions.

Among the 114 mutations of *MSH2* is the A→T change at nt942+3 of the splice donor site of exon 5, which appears world wide. It has been reported in over 20 patients or families not counting the ones in Newfoundland.[4 5] As it appears to occur in virtually all populations that have been extensively studied so far,[6–9] we wished to exclude the possibility that present day mutation carriers might descend from a distant common ancestor. The alternative hypothesis that the mutation arises de novo with a relatively high frequency appeared more likely in view of our previous studies.[4 5] If this hypothesis could be proven correct, it would suggest the existence of a hitherto unexplored mechanism that predisposes to this particular change.

Our results lend full support to the hypothesis that the mutation has arisen de novo on multiple occasions. This does not by any means invalidate our previous findings in Newfoundland patients. In the Newfoundland study, eight of 11 families share a mutation associated haplotype of some 10 cM in length. Such an extensive shared haplotype is fully compatible with a single major founding event in Newfoundland, the mutation having been introduced by an early settler some time after 1610, that is, less than 15 to 20 generations ago. In contrast, we found no evidence of a shared haplotype in English or Italian families even with a battery of intragenic SNP and microsatellite markers that we calculate would not sustain more than a few historical recombinations in 500 to 2000 generations. Moreover, three Asian families also did not show evidence of haplotype sharing. It should perhaps be mentioned that even with several intragenic markers studied in 10 different kindreds, formal exclusion of haplotype sharing in occasional kindreds cannot be claimed, because many marker alleles in mutation associated haplotypes are also the most common alleles in control haplotypes. For example, focusing on the three intragenic SNPs, the haplotype 111 on the disease chromosome is shared by as many as six families, whereas other haplotypes are seen in the remaining four families. However, in control chromosomes, the 111 haplotype is seen 13 times whereas other haplotypes occur nine times (table 5). Thus, a p value of 0.38 suggests no difference between these two distributions. This analysis therefore strongly suggests the absence of a shared

ancestral haplotype, but obviously cannot prove it. Nevertheless, we consider our data to be strong because we were able to analyse not only these three intragenic SNPs, which have only moderately informative allele frequencies, but also six microsatellite polymorphisms which are highly informative. All our results suggest that the mutation recurs de novo.

What is the mechanism that strongly predisposes the A nucleotide in position 942+3 to become a T instead? We assume that the change seen in germline tissue occurs in meiosis, which makes it relatively unlikely to be the result of environmental influence other than ionising radiation. We have no reason to implicate ionising radiation; however, to exclude such an effect may be difficult. Instead, we are inclined to consider whether the DNA sequence itself in the immediate vicinity of the affected nucleotide might contribute to the risk. That a high risk is indeed present not only in meiosis but also in mitosis is evident from one pertinent previous observation. One of us found the same A→T change as a somatic mutation ("second hit") in an endometrial cancer that developed in a germline carrier of another *MSH2* mutation in an HNPCC family (M Miyaki, unpublished observations). Of note, somatic mutations of *MSH2* have been described relatively rarely,[3] so a high relative frequency of the A→T change as a somatic event is a possibility. This would considerably strengthen the likelihood of it being because of a sequence peculiarity that affects meiosis as well as mitosis.

The intron between exons 5 and 6 has not been fully sequenced, but the BAT-26 locus contains a 26 adenine repeat beginning with the third nucleotide of the intron, that is, the A that is replaced by a T in the mutation. Thus, the A→T mutation occurs at the first position in this highly mutable sequence. BAT-26 was initially thought to be quasi-monomorphic ($(A)_{26}$) in the population, displaying only extremely rare alleles being deleted for one ($(A)_{25}$) or two ($(A)_{24}$) nucleotides.[29] However, in Africans and African-Americans, it displays outright widespread polymorphism with alleles containing repeat lengths as short as ($(A)_{20}$) to ($(A)_{10}$).[30 31] A distinctive feature of this repeat is its extreme susceptibility to deletion in mismatch repair deficient tumours.[33] It is virtually always deleted in tumours that show a high degree of microsatellite instability and is, therefore, widely used as a marker for mismatch repair deficiency.[32 33] Mutation mechanisms resulting from DNA replication errors occur by both base mispairing and strand slippage that leads to either base pair substitution or insertion/deletion respectively. This study indicates that it is a de novo mutational hot spot for base pair substitution. In a somewhat analogous case, Laken *et al*[34] have reported a mutation (T to A at APC nucleotide 3920) found in 6% of Ashkenazi Jews and about 28% of Ashkenazim with a family history of CRC. This mutation creates an $(A)_8$ repeat that constitutes a small hypermutable region of the gene, indirectly causing cancer predisposition.

Do mutational hotspots occur in *MLH1* as well? Among 161 different mutations described in the database, two stand out as being highly recurrent; however, both have so far been seen exclusively in ethnic Finns. Extensive haplotype analyses have shown that both represent ancient founder mutations enriched in the Finnish population.[24 35] Two further mutations have been seen more than just a few times. A change of C→T in codon 117 of exon 4 has been seen at least seven times, and a deletion of AAG in codon 616 of exon 16 has been reported in some 13 families world wide. If these turn out to arise recurrently de novo, they may be additional examples, albeit not as prevalent, of true mutational hotspots of unknown causation.

The overall significance of recurring mutations is at least twofold.[36] First, in populations where certain mutations are so enriched (by a founder affect) that they account for a high proportion of all mutations, they have obvious diagnostic implications. This is the case with the present mutation in Newfoundland (27% of all HNPCC) and the two prevalent *MLH1* mutations in Finland (>50% of all HNPCC). Second, mutations such as the present one that arise spontaneously de novo probably result from a predisposition of either environmental or genetic nature or both. Currently these cannot be distinguished, but by eventually elucidating the mechanisms in detail, clues to their prevention may emerge.

We thank Dr Bo Yuan for analysis of the *MSH2* sequence and Dr Natalia Pellagata for assistance with molecular genetics analysis. This study was supported in part by grant P30CA16058, the National Cancer Institute, Bethesda, Maryland. PR was supported by grants from the Italian Association and Foundation for Cancer Research (AIRC/FIRC). FAW was supported in part by NIH GM58934. AdlC was supported by grants CT940676 from the European Union and CA67941 from the National Institutes of Health.

1 Peltomaki P, Aaltonen LA, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Green JS, Jass JR, Weber JL, Leach FS, Petersen GM, Hamilton SR, de la Chapelle A, Vogelstein B. Genetic mapping of a locus predisposing to human colorectal cancer. *Science* 1993;**260**:810-12.
2 Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M, Guan XY, Zhang Z, Meltzer PS, Yu JW, Kao FT, Chen DJ, Cerosaletti KM, Fournier REK, Todd S, Lewis T, Leach RJ, Naylor SL, Weissenbach J, Mecklin JP, Jarvinen H, Petersen GM, Hamilton SR, Green J, Jass J, Watson P, Lynch HT, Trent JM, de la Chapelle A, Kinzler KW, Vogelstein B. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 1993;**75**:1215-25.
3 Liu B, Parsons RE, Hamilton SR, Petersen GM, Lynch HT, Watson P, Markowitz S, Willson JK, Green J, de la Chapelle A, Kinzler KW, Vogelstein B. h*MSH2* mutations in hereditary nonpolyposis colorectal cancer kindreds. *Cancer Res* 1994;**54**:4590-4.
4 Froggatt NJ, Joyce JA, Davies R, Evans DGR, Ponder BA, Barton DE, Maher ER. A frequent h*MSH2* mutation in hereditary non-polyposis colon cancer syndrome. *Lancet* 1995;**345**:727.
5 Froggatt NJ, Green J, Brassett C, Evans DG, Bishop DT, Kolodner R, Maher ER. A common *MSH2* mutation in English and North American HNPCC families: origin, phenotypic expression, and sex specific differences in colorectal cancer. *J Med Genet* 1999;**36**:97-102.
6 Viel A, Genuardi M, Capozzi E, Leonardi F, Bellacosa A, Paravatou-Petsotas M, Pomponi MG, Fornasarig M, Percesepe A, Roncucci L, Tamassia MG, Benatti P, Ponz de Leon M, Valenti A, Covino M, Anti M, Foletto M, Boiocchi M, Neri G. Characterization of *MSH2* and *MLH1* mutations in Italian families with hereditary nonpolyposis colorectal cancer. *Genes Chrom Cancer* 1997;**18**:8-18.
7 Miyaki M, Konishi M, Muraoka M, Kikuchi-Yanoshita R, Tanaka K, Iwama T, Mori T, Koike M, Ushio K, Chiba M, Nomizu S, Utsunomiya J. Germ line mutations of h*MSH2* and h*MLH1* genes in Japanese families with hereditary nonpolyposis colorectal cancer (HNPCC): usefulness of DNA analysis for screening and diagnosis of HNPCC patients. *J Mol Med* 1995;**73**:515-20.

8 Danish HNPCC Register. Quoted in the HNPCC database (www.nfdht.nl).
9 Liu B, Parsons R, Papadopoulos N, Nicolaides NC, Lynch HT, Watson P, Jass JR, Dunlop M, Wyllie A, Peltomäki P, de la Chapelle A, Hamilton SR, Vogelstein B, Kinzler KW: Analysis of mismatch repair genes in hereditary non-polyposis colorectal cancer patients. *Nat Med* 1996;**2**:169-74.
10 Chan TS, Yuen ST, Chung LP, Ho JWC, Kwan KYM, Chan ASY, Ho JCY, Leung SY, Wyllie AH. Frequent microsatellite instability and mismatch repair gene mutations in young Chinese patients with colorectal cancer. *J Natl Cancer Inst* 1999;**91**:1221-6.
11 Wang Q, Lasset C, Desseigne F, Saurin JC, Maugard C, Navarro C, Ruano E, Descos L, Trillet-Lenoir V, Bosset JF, Puisieux A. Prevalence of germline mutations of h*MLH1*, h*MSH2*, h*PMS1*, h*PMS2*, and h*MSH6* genes in 75 French kindreds with nonpolyposis colorectal cancer. *Hum Genet* 1999;**105**:79-85.
12 Pensotti V, Radice P, Presciuttini S, Calistri D, Gazzoli I, Perez APG, Mondini P, Buosanti G, Sala P, Rossetti C, Ranzini GN, Bertario L, Pierotti MA. Mean age of tumor onset in hereditary nonpolyposis colorectal cancer (HNPCC) families correlates with the presence of mutations in DNA mismatch repair genes. *Genes Chrom Cancer* 1997;**19**:135-42.
13 Bai YQ, Akiyama Y Nagasaki H, Lu SL, Arai T, Morisaki T, Kitamura M, Muto A, Nagashima M, Nomizu T, Iwama T, Itoh H, Baba S, Iwai T, Yuasa Y. Predominant germ-line mutation of the h*MSH2* gene in Japanese hereditary nonpolyposis colorectal cancer kindreds. *Int J Cancer* 1999; **82**:512-15.
14 Fidalgo P, Almeida MR, West S, Gaspar C, Maia L, Wijnen J, Albuquerque C, Curtis A, Cravo M, Fodde R, Leitao CN, Burn J. Detection of mutations in mismatch repair genes in Portugese families with hereditary non-polyposis colorectal cancer (HNPCC) by a multi-method approach. *Eur J Hum Genet* 2000;**8**:49-53.
15 Kan YW, Dozy AM. Antenatal diagnosis of sickle-cell anaemia by DNA analysis of amniotic-fluid cells. *Lancet* 1978;**2**:910-12.
16 Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek M Jr, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo, MP, Garnerone S, Restagano G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, Ferec C, de Arce M, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L. The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat Genet* 1994;**7**:169-75.
17 Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 1992;**2**:204-11.
18 Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, Breakefield X, Bressman S. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* 1995;**9**:152-9.
19 Chadwick RB, Conrad MP, McGinnis MD, Johnston-Dow L, Spurgeon SL, Kronick MN. Heterozygote and mutation detection by direct automated fluorescent DNA sequencing using a mutant Taq DNA polymerase. *Biotechniques* 1996;**20**:676-83.
20 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;**58**:1347-63.
21 Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987;**51**:227-33.
22 Gao X, Wright FA. Nonparametric disequilibrium mapping when haplotypes are available. *Am J Hum Genet Suppl* 1999;**65**:A250.
23 Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* 1995;**56**: 777-87.
24 de la Chapelle A, Wright FA. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc Natl Acad Sci USA* 1998;**95**:12416-23.
25 Strachan T, Read AP. *Human molecular genetics*. Chapter 10. New York: Wiley, 1997.
26 Wijnen J, van der Klift H, Vasen H, Khan PM, Menko F, Tops C, Meijers Heijboer H, Lindhout D, Moller P, Fodde R. *MSH2* genomic deletions are a frequent cause of HNPCC. *Nat Genet* 1998;**20**:326-8.
27 Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, Lynch HT, Chadwick RB, de la Chapelle A, Markowitz S, Laken SJ, Lengauer C, Kinzler KW, Vogelstein B. Conversion of diploidy to haploidy. *Nature* 2000;**403**:723-4.
28 Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, Harkonen N, Julkunen R, Kangas E, Ojala S, Tulikoura J, Valkamo E, Jarvinen H, Mecklin JP, Aaltonen LA, de la Chapelle A. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol* (in press).
29 Zhou XP, Hoang JM, Cottu P, Thomas G, Hamelin R. Allelic profiles of mononucleotide repeat microsatellites in control individuals and in colorectal tumors with and without replication errors. *Oncogene* 1997;**15**:1713-18.
30 Pyatt R, Chadwick RB, Johnson CK, Adebamowo C, de la Chapelle A, Prior TW. Polymorphic variation at the BAT-25 and BAT-26 loci in individuals of African origin. Implications for microsatellite instability testing. *Am J Pathol* 1999;**155**:349-53.
31 Samowitz WS, Slattery ML, Potter JD, Leppert MF. BAT-26 and BAT-40 instability in colorectal adenomas and carcinomas and germline polymorphisms. *Am J Pathol* 1999;**154**:1637-41.
32 Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* 1998;**58**:5248-57.
33 de la Chapelle A. Testing tumors for microsatellite instability. *Eur J Hum Genet* 1999;**7**:407-8.
34 Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in *APC*. *Nat Genet* 1997;**17**:79-83.
35 Moisio AL, Sistonen P, Weissenbach J, de la Chapelle A, Peltomaki P. Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. *Am J Hum Genet* 1996;**59**:1243-51.
36 Lynch HT, de la Chapelle A. Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* 1999;**36**:801-18.