Origins and Prevalence of the American Founder Mutation of MSH2

Mark Clendenning, 1 Mark E. Baze, 2 Shuying Sun, 3 Kyle Walsh, 2 Sandya Liyanarachchi, 1 Dan Fix, 1 Victoria Schunemann, 2 Ilene Comeras, 2 Molly Deacon, 4 Jane F. Lynch, 1 Gordon Gong, 4 Brittany C. Thomas, 5 Stephen N. Thibodeau, 5 Henry T. Lynch, 4 Heather Hampel, 1 and Albert de la Chapelle 1

1Human Cancer Genetics Program and 2Department of Internal Medicine, Comprehensive Cancer Center and 3Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio; Department of Preventive Medicine, Creighton University School of Medicine, Omaha, Nebraska; and 4Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota

Abstract

Large germline deletions within the mismatch repair gene MSH2 account for a significant proportion (up to 20%) of all deleterious mutations of this gene which are associated with Lynch syndrome. An exons 1 to 6 deletion of MSH2, originally reported in nine families, has been associated with a founding event within the United States, which genealogic studies had previously dated to 1727, and the number of present day carriers was estimated to be 18,981. Here, we report the development of a robust multiplex PCR which has assisted in the detection of 32 new families who carry the MSH2 American Founder Mutation (AFM). By offering testing to family members, 126 carriers of the AFM have been identified. Extensive genealogic studies have connected 27 of the 41 AFM families into seven extended pedigrees. These extended families have been traced back to around the 18th century without any evidence of further convergence between them. Characterization of the genomic sequence flanking the deletion and the identification of a common disease haplotype of between 0.6 and 2.3 Mb in all probands provides evidence for a common ancestor between these extended families. The DMLE+2.2 software predicts an age of ~500 years (95% confidence interval, 425–625) for this mutation. Taken together, these data are suggestive of an earlier founding event than was first thought, which likely occurred in a European or a Native American population. The consequences of this finding would be that the AFM is significantly more frequent in the United States than was previously predicted. [Cancer Res 2008;68(7):2145–53]

Introduction

Lynch syndrome, which is also known as hereditary nonpolyposis colorectal cancer, is estimated to account for approximately one third of all heritable colon cancers and 2.8% of all colon cancers in the Western world (1). 6 This amounts to ~4,500 cases a year in the United States alone (2). It is currently accepted that Lynch syndrome arises as a result of mutations in mismatch repair genes, giving the syndrome a dominant inheritance pattern. Four mismatch repair genes, MLH1, MSH2, PMS2, and MSH6, are typically implicated in the transmission of Lynch syndrome, with the majority of mutations having been identified in either MLH1 or MSH2.

A broad spectrum of mutations have been identified throughout these genes, but due to the detection capabilities for large rearrangements being relatively poor until recently, the majority of mutations identified to date have been point mutations or small insertion/deletions. With the development of improved techniques, such as multiplex ligation-dependent probe amplification (MLPA; ref. 3) and semiquantitative PCR (4), the number of large genomic rearrangements identified has increased considerably. One such mutation is a deletion of exons 1 to 6 of MSH2, which was initially identified in nine families from the United States (5, 6). The MSH2 American Founder Mutation (AFM) was characterized by long-range PCR and sequencing of the breakpoints, which allowed for the separation of this deletion from that of other similar, but distinct, exons 1 to 6 deletions. Subsequent genealogic studies had linked three of the nine families and traced them through 12 generations to a putative common ancestor who migrated to the United States from Germany in the early 18th century (7). Based on this information, it was calculated that 18,981 [95% confidence intervals (95% CI), 6,038–34,466] Americans carry the AFM mutation (8).

In this study, we sought to gain further insight into the origin, occurrence, and spread of the MSH2 AFM in the United States. We developed a robust multiplex PCR which can be performed using standard conditions and thus allows for the accurate identification of AFM carriers. Using this detection method, we identified a further 32 families who carry the AFM. Haplotypes flanking the mutation were characterized in 29 of these new families, as well as four of the original families (throughout this article, we refer to the original families, families A–I, as families 1–9, respectively). The use of an intraallelic coalescent model generated an estimate for the age of the mutation, which may have considerable implications for the predicted origin and prevalence of the MSH2 AFM.

Materials and Methods

Patient samples. This study was advertised to cancer genetic counselors in the United States through their e-mail listserv, their newsletter “Perspectives in Genetic Counseling,” and their research directory. In addition, four families were identified through our own population-based genetic testing study for Lynch syndrome (9). 6 Free predictive genetic testing was offered to at-risk members of any family with an exons 1 to 6

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Albert de la Chapelle, Human Cancer Genetics, Room 804, Biomedical Research Tower, 460 West 12th Avenue, Columbus, OH 43210. Phone: 614-688–4781; Fax: 614-688–4772; E-mail: albert.delachapelle@osumc.edu.

©2008 American Association for Cancer Research. doi:10.1158/0008-5472.CAN-07-6599

Diploid-to-haploid conversion. Haploid-converted clones from patient 8 were created using the conversion technology of Yan et al. (10). In brief, human lymphoblastoid cell lines were electrofused with a specifically designed mouse cell line (E2). Unfused mouse parental cells were negatively selected by sodium hypoxanthine, aminopterin, and thymidine (HAT), and unfused human lymphocytes were negatively selected by Genetecin. Hybrid cells were maintained in DMEM, including 10% fetal bovine serum, 0.5 mg/mL Genetecin, 1× HAT, and penicillin-streptomycin. To ensure that each clone contained one of the two homologues of chromosome 2, clones were typed at two polymorphic microsatellite loci (D2S2952 and D2S434).

Genotype analysis. Twelve microsatellite and seven single-nucleotide polymorphism (SNP) markers (Supplementary Table S1 and Supplementary Fig. S1) were used to obtain a haplotype spanning ~12.5 Mb across the MSH2 locus. Markers were typed in AFM-proband-positive and a set of 118 Caucasian control DNAs (Coriell Institute). SNPs were typed using standard PCR and sequencing protocols. For the microsatellite analysis, the reverse primer had an M13 tail which, in combination with a FAM-labeled M13 oligo, could be sized using an ABI7000. Each 25-μL PCR reaction contained 12.5 μL of AmpliTaq Gold master mix (Applied Biosystems), 25 ng of genomic DNA, 10 pmol of forward primer, 2 pmol of tagged reverse primer, and 10 pmol of the FAM-labeled M13 primer. Reactions were multiplexed when possible and cycled using the following profile: 96°C for 10 min, 50 cycles of 96°C for 30 s, 60°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 10 min.

Estimating allele age. To make an estimate for the age of the AFM, we took two approaches, both of which use the haplotype data generated from affected individuals and data obtained from a set of unaffected controls. The first method, described by Risch et al. (11), gives a separate age estimate for each marker using the following calculation: age (in generations) = ln(LD) / ln(1 − θ), wherein LD is the linkage disequilibrium index (12) and θ is the recombination fraction between the marker of interest and the mutation. Genetic distances were calculated using a conversion factor of 1.1 Mb/cM, which was obtained from a comparison of Centimorgan and megabase values for the 2p22-2p16 region (based on deCODE mapping data; ref. 13). The second method uses an intraallelic coalescent model to assess the linkage disequilibrium across the marker set as a whole. The analysis is performed using the DMLE+2.2 software (14).

In addition to the genotype data, marker locations, population growth rates, and an estimate for the proportion of disease bearing chromosomes being analyzed are used by the software. To estimate the proportion of disease-bearing chromosomes being studied, we used the following data, 6% lifetime risk for CRC (American Cancer Society),11 2.8% of CRCs are Lynch syndrome,6 and 6.8% of Lynch syndrome cases are due to the AFM.6 For the population growth rate, we took data from the U.S. census, which records population figures back to 1790.12 The census records show that the population growth rates have decreased with time; so to account for this, we used a growth rate across the whole time period [1.65-fold per generation (fpg), where a generation is taken as 25 y] along with two extreme growth rates (one from the 19th century, 1.96 fpg, and one from the 20th century, 1.38 fpg) which will be representative of our outer CIs.

Results

Identification of AFM carriers. As part of a large study of all newly diagnosed colorectal and endometrial cancer patients from the Columbus metropolitan region of Ohio (9),6 microsatellite instability (MSI) status was determined in 1,566 CRC patients. All 306 MSI-positive cases were screened via MLPA for large-scale alterations within the MSH2 gene. Eight of these samples gave aberrant results typical of a deletion or insertion in MSH2, four of

---

8 Usually a combination of information found at either www.ancestry.com or www.familysearch.org and from privately known/researched information.
9 www.mrc-holland.com
which suggested the deletion of exons 1 to 6 of MSH2. In addition, 32 patients from throughout the United States were identified as having an exons 1 to 6 deletion of MSH2 at an outside commercial laboratory (Mayo Medical Laboratories and Myriad Genetic Laboratories, Inc.) and were brought to the attention of this study. Due to the sequence composition of this region of the genome (i.e., excessive numbers of Alu-like repeats; refs. 4, 15), it is essential to confirm the exact breakpoint sequence of this deletion, as it is known that other exons 1 to 6 deletions occur as a result of recombination between alternate repeat elements or even at different locations within the same repeat elements (16, 17). To this end, each of the 36 samples were amplified using the new breakpoint-specific primers, and any break-specific products were sequenced to confirm the precise location of the break. All four of our population-based cases and 28 of the 32 outside cases were found to carry the same AFM-specific mutation (Fig. 1).

As motivation for assisting us with our genealogic studies, we offered free predictive testing to any at-risk individuals from our mutation-positive families. These relatives were tested (often in large family groups) after receiving thorough genetic counseling, which is a key component of the testing process. Through this incentive, we received samples from 171 relatives of these 41 AFM probands for mutation screening; 85 (49.7%) of which were shown to carry the AFM.

For each proband who carried the AFM, a geographic location was recorded and plotted onto a map of the United States (Fig. 2). From these data, we were able to see that present day families were primarily located within three States; Kentucky, Ohio, and Texas. To address the issue of ascertainment bias being an apparent cause for the prevalence within Ohio and Kentucky, we sought to obtain additional data on AFM carriers from two of the nation’s most prominent genetics laboratories, the Mayo Clinic13 and Myriad Genetics.14 In keeping with HIPAA guidelines, both laboratories were unable to provide certain patient-specific information; however, they were able to report that the incidences of the AFM among Lynch syndrome patients in their cohorts was 4.3% and 7%, respectively, as well as the geographic locations for their AFM-positive probands (Fig. 2).

Throughout the course of our study, we have actively sought to identify a carrier of the AFM outside of the United States, with efforts focusing on European countries as the most likely source of any founding individuals. To date, we have screened three samples from France (two French and one Belgian), two from Sweden, three from Italy, two from Poland, six from Germany (five Germans and one Armenian), one from Scotland, and four from Ireland (three Northern and one Republic) known to have an MSH2 exons 1 to 6 deletion, and none of them has proved to have the AFM-specific breakpoints.

Genotype characterization. To link the probands genetically and rule out the possibility that the mutation was occurring in the affected families de novo, we typed a set of polymorphic microsatellite markers and SNPs which flanked the mutation site. A panel of 12 microsatellites which flanked the mutation site and had an average interval of 956 kb (range, 120 kb–2 Mb) was chosen. We optimized microsatellite selection based on three criteria: (a) the repeat had to be highly polymorphic to reduce likelihood of similarity by chance, (b) the microsatellite and its flanking sequence had to be unique within the genome, and (c) the repeat was not located close to or within a more complex repeat, such as an Alu, which can often introduce additional size variation due to poly(A) tracts and can cause problems with primer binding due to

---

13 http://www.mayoclinic.com
14 http://www.myriad.com
their excessive prevalence. The use of a haploid cell line from patient 8 enabled us to generate an inferred haplotype for 34 samples (29 from this study and 5 from the original study; Fig. 3) and showed that all patients shared a common disease haplotype. Seven informative SNPs were typed at selected regions to establish if a specific microsatellite was different from the consensus because of a recombination event rather than a mutation. Twenty-one of the patients have disease-specific haplotypes in excess of 4.99 Mb; however, the core haplotype shared by all 34 of the patients is between 0.59 Mb (Clen32-Clen30) and 2.26 Mb (rs10495934-Clen25). To ensure that allele match was not a chance event, a panel of 118 control DNAs were typed with the 19 markers to determine the degree of variation for each marker (Fig. 3).

**Genealogic studies.** Excluding one proband (20) whose father (from whom the mutation was inherited) was adopted, there were 40 families (31 families identified in this study and 9 families from the original study) with the AFM available for genealogic studies. Using the techniques described in Materials and Methods, 27 of these ostensibly unrelated families were linked into seven pedigrees, which we refer to herein as the “subfounder” groups, denoted A to G (Figs. 4 and 5). We have been unable, thus far, to link the seven subfounder groups to a single common ancestor to elucidate the true founder of this mutation in the United States. However, the oldest of the seven subfounder pedigrees (B) seems to trace back to a husband and wife who were born in the United States in about 1700, with the other pedigrees all tracing back to couples born in the mid to late 1700s. These individuals all seem to have been born, where known, exclusively in either Virginia or North Carolina. In the case of subfounder group A, which contains highly intermarried lines descending primarily from two early pioneering families, there are three different progenitor couples to which the eight current day families (1, 29, 30, 31, 33, 35, 39, and 40) can be linked, and as such, there are five different individuals who could have introduced the AFM into this subfounder group (Fig. 4A, the five potential founders are marked with an asterisk).

From the attainable data, we can trace six of the seven (A, B, C, D, E, and G) subfounder groups back as having migrated through Kentucky (where some families remained), with data showing that many of the “founding” couples were married in relatively close geographic proximity to one another (A, southwestern Virginia ca.

![Figure 2. Map of the United States depicting the location of patients, representing each of 76 families (at the time of diagnosis) and subfounders. Patients described in this study are shown as squares (numbered 1 through 41), patients identified through Myriad Genetics Laboratories, Inc., are shown as triangles, and patients identified through the Mayo Clinic are shown as circles. The location of common ancestors for the subfounder pedigrees are displayed as stars, wherein A, C, D, and E correspond to marriage locations of the common ancestors, and B, F, and G correspond to the regions that the common ancestors were born in. Due to the complexity of pedigrees A and B, there is some ambiguity for these subfounder locations.](image-url)
**Figure 3.** Genotype data spanning the MSH2 locus of 36 DNAs. The 36 samples comprise 34 patients and 2 haploid clones. Disease associated genotypes are highlighted in gray, with the disease allele depicted in red. The size of the minimum and maximum conserved haplotypes along with the frequency of the disease associated allele among a panel of control DNAs are shown at the bottom. The patients have been ordered according to their subfounder groups, and where possible, we have displayed family specific haplotypes, which differ from the disease haplotype, in a colored font. Based on these data it is likely that patient 34 is part of subfounder group A, due to the presence of a family specific haplotype at the 3' end of the gene (blue font).

1778; C, western North Carolina ca.1820; D, northeastern Tennessee.\(^{15}\) ca. 1785; E, northeastern Tennessee.\(^{16}\) ca. 1789; Fig. 2). Although a mixture of European ancestry is evident among these subfounder groups, the predominant ancestry seems to be from the British Isles, primarily Scottish. Various families in most of the subfounder groups also have reported Native American ancestry, specifically Cherokee, but thus far this seems verifiable in only three or four families. These Native American links are certainly possible because the geographic placement of the progenitor couples, as given above, is either just inside or very close to the known Cherokee territorial boundaries of the time.\(^{11}\)

Subfounder group D includes families 6 and 7 from the original AFM study.\(^{7}\) These two families connect to three newly identified AFM families (24, 25, and 28) through the wife of an individual who was originally thought to be an obligate carrier of the mutation due to him being an apparent link between three of the original families (Fig. 5A). Given the certainty of the new genealogy connections, the genealogy of family 5 (the only other family that was linked to families 6 and 7) was scrutinized. It was discovered that there was an error in the original genealogy due to two men sharing the same name. Given the locations of these men at the time (one in Illinois and the other in Alabama), it is clear that they were two different individuals and that family 5 does not link to families 6 and 7, as described in the original paper. This also proved that the true AFM “founder family” could not be the family who immigrated to the United States in 1727 from Germany as reported (7).

**Estimate of allele age.** It has been shown that the rate of recombination and population frequency of flanking marker alleles can be used to estimate the age of a mutation. In this study, we used two established methods: a single marker approach (11) and a combined marker approach (14). Using a marker-by-marker approach, we generated age estimates at all loci with respect to the AFM mutation (Supplementary Table S2; refs. 18, 19). Each marker estimate is generated independently from all the others, based on its location with respect to the mutation (recombination fraction) and the frequency with which the disease allele is found within the normal population. When all markers are considered, we get a

---

\(^{15}\) Present-day northeastern Tennessee was part of North Carolina during the described period.

\(^{16}\) http://www.libs.uga.edu/darchive/hargrett/maps/maps.html
Figure 4. Parsimonious pedigrees showing the genealogically linked families. Pedigrees show direct blood line links only; siblings and spouses are included when they seem to carry the mutation or if they are equally likely to be a founder lineage, respectively (full pedigree data can be accessed online, http://internalmedicine.osu.edu/genetics/990.cfm). Individuals are labeled with their initials and year of birth, with the presence of a Lynch syndrome–associated cancer (endometrial, uterine, ovarian, pancreatic, ureter and renal pelvis, biliary tract, brain, small bowel, sebaceous adenomas, and keratoacanthoma) being depicted as a solid black symbol when known. Pedigrees are drawn to an approximate scale of 25 y/generation as displayed at the left hand side of each pedigree. Asterisks have been placed at the top right corner of the five individuals who could all have introduced the mutation into subfounder group A.
broad estimate of 4 to 98 generations (100–2,450 years), with 13 of 15 markers giving a range of 4 to 26 generations (100–650 years). The two markers which define the outer limits of the fully conserved haplotype (rs10495934 and Clen25) give an estimate of 23 generations (575 years) and 19 generations (475 years), respectively.

The second method, which considers the linkage disequilibrium across all markers in a single analysis, is implemented in the

![Diagram A](image1)

![Diagram B](image2)

![Diagram C](image3)

![Diagram D](image4)

Figure 5. Parsimonious pedigrees showing the genealogically linked families, continued.
Discussion

Given the fact that mutations in the mismatch repair gene MSH2 account for a significant proportion (1–2%) of all cases of colorectal cancer, it is important to assess the risk that can be attributed to specific, frequently occurring, mutations. In this study, we have taken both molecular-based and genealogic-based approaches to get a better understanding of the prevalence and spread of the AFM of MSH2. Screening a large cohort of Lynch syndrome families added an additional 32 ostensibly unrelated families to the nine which had been identified previously (6). Genotype analysis in 34 of these families showed a conserved haplotype of some 5 Mb in most cases and a much smaller haplotype (0.59–2.67 Mb) in a single case (34), which allowed for an age calculation of ~500 years for the AFM. Whereas haplotyping proves that these modern-day AFM carriers are all distantly related to one another, the oldest genealogic links are to individuals born during the early 18th century in Virginia. As it becomes ever more difficult to extend these genealogies before 1700, due to the sparsity of recorded data from these times, it is apparent that we may never know who the original AFM founding individual was. This MSH2 AFM has some similarities to a recently described AFM in the APC gene which was brought to the United States from England around 1630 and affects two large U.S. kindreds (20).

Our new genealogic findings and haplotyping data do however provide very strong evidence against the hypothesis that the AFM arrived in the United States as a single event in the early 18th century. The records linking 27 of the AFM families into seven extended families each with a common ancestor being born between 1700 and the early 1800s would make it seem very unlikely that they could all converge on each other in such a period of time that would be consistent with a single common ancestor having arrived in the United States during the early 18th century, a time of significant European immigrations. With this in mind, we are left with two hypotheses, both of which are supported by the mutation age calculations. Either, the subfounder families were in the United States for several generations before that of the current subfounder common ancestors, which would allow for the possibility that the mutation was either brought into the United States by a single European immigrant during an earlier period or that the mutation had been introduced into these European lines by a Native American individual during this early colonial period. Or, the mutation originated within Europe several generations before it arrived on the shores of the United States, probably through several individuals.

The lack of evidence for the AFM outside of the United States is perhaps more supportive of the first hypothesis; however, there also lies the possibility that most of the possible carriers of the mutation in Europe may have emigrated or that many failed to pass on the mutation to subsequent generations and so we might predict a significant reduction of cases, or even an absence of cases in the founding country. In addition, our search for the AFM in Europe has by no means been exhaustive, so it is still possible that there are AFM mutation carriers in Europe that we have just not been able to identify yet. We have speculated, based on our genealogic studies, that Scotland in particular is a potential European source for the mutation; however, we have only been able to identify a single candidate (individual with a known exons 1–6 deletion of MSH2) from this region, who was shown not to carry the AFM. In addition, samples we have screened from an ancestrally related population (Ireland) were also negative for the mutation, and to the best of our knowledge, it has not been described elsewhere in the British Isles. Until evidence in support of one of these two hypotheses is obtained, whether it is the identification of a single common ancestor within the United States or the presence of a case of the AFM in another country,
we cannot make firm conclusions as to the origins of this mutation. Nevertheless our current data show that the AFM is significantly more prevalent within the United States than was previously thought, and is clearly a prominent cause of Lynch syndrome, particularly in Ohio, Kentucky, and Texas. This is an important public health issue because there is ample evidence that individuals who are identified as having Lynch syndrome can benefit from highly targeted cancer control measures (21–23).

With this in mind, we would promote the introduction of a single site diagnostic PCR, such as the one described here, as a first line of molecular screening in patients whose tumors stain negatively for MSH2 at the protein level. Such a diagnostic test would not only reduce the cost of genetic screening in a significant proportion of Lynch syndrome cases, but it would also allow for a far more rapid diagnosis than techniques, such as MLPA, and complete gene sequencing.

Acknowledgments

Received 12/10/2007; revised 1/15/2008; accepted 1/16/2008.

Grant support: NIH grants CA67941 and CA16658 and NSF grant 0112050.

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

We thank Thierry Frebourg, Amika Lindblom, Alessandra Viel, Jan Lubinski, Elisabeth Mangel, Timm Goecke, Patrick Morrison, Andrew Green, Malcolm Dunlop, and Rebecca Bartelson for the provision and screening of MSH2 exons 1 to 6 deletion cases; Thomas Prior and the Ohio State University Molecular Pathology Laboratory for confirming the breakpoints in all the probands and for providing testing to the at-risk family members; Douglas Crews for his contributions to the genealogic studies; Richard Wenstrup, Lynn Anne Burbridge, and Cynthia Frye of Myriad Genetic Laboratories, Inc., for the provision of epidemiologic data; and the families who participated in this research.

References

Origins and Prevalence of the American Founder Mutation of MSH2

Mark Clendenning, Mark E. Baze, Shuying Sun, et al.


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/68/7/2145

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2008/03/28/68.7.2145.DC1

Cited articles
This article cites 23 articles, 4 of which you can access for free at:
http://cancerres.aacrjournals.org/content/68/7/2145.full.html#ref-list-1

Citing articles
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
/content/68/7/2145.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.