"New Haplotype and Genealogical Data Give Important Implications for the Origins and Prevalence of the American Founder Mutation of *MSH2*"

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation
with research distinction in Molecular Genetics in the undergraduate
College of Arts and Sciences at The Ohio State University

by

Kyle M. Walsh

The Ohio State University
May, 2007

Project Advisor: Professor Albert de la Chapelle MD, PhD
Department of Molecular Virology, Immunology and Medical Genetics

Bachelor's Examination Committee:

Dr. A. de la Chapelle, Department of M.V.I.M.G.
Dr. A. Simcox, Department of Molecular Genetics
Dr. D. Crews, Department of Anthropology
H. Hampel, C.G.C., Department of Internal Medicine

Approved by:

_____

Dr. Albert de la Chapelle, Advisor
Department of Molecular Virology, Immunology and Medical Genetics

Acknowledgements:

I would like to thank my P.I., Dr. Albert de la Chapelle, for the opportunity to work as a member of his world-class laboratory which contains a host of bright and talented researchers. These researchers are very deserving of my gratitude including, first and foremost, my post-doctoral advisor Dr. Mark Clendenning whose guidance has been greatly appreciated and often much needed. I would also like to thank Heather Hampel, who has always been my primary source of direction on this project. Additionally, Dr. Douglas E. Crews has been instrumental in helping me to bridge the gap between my two disciplines, molecular genetics and anthropology.

As this work has always been a team project spearheaded by the individuals listed above and myself, I need to thank the other team members who have contributed to this work, including: Dan Fix, Mark Baze, Ilene Comeras and Sandya Liyanarachchi.

For financial support I must thank the de la Chapelle lab, The Ohio State University College of Arts and Sciences, The OSU College of Biological Science, and the Undergraduate Student Government.

Finally, I appreciate the astounding willingness of the genetic counseling patients to provide me with family history information and help in the genealogical research. One of these patients, who made this work the primary use of his free time, deserves immense credit for his dedication to this study.

Abstract:

Lynch syndrome, which is also known as hereditary nonpolyposis colorectal cancer (HNPCC), is estimated to account for approximately half of all heritable colon cancers and perhaps upward of 3% of all colon cancers in the western world. This amounts to approximately 4500 cases a year in the US alone.[1] Lynch syndrome arises as a result of mutations in genes of the mismatch repair pathway, conferring an autosomal dominant cancer predisposition with an ~80% lifetime risk of HNPCC-associated cancers.[1] Four mismatch repair (MMR) genes, *MLH1*, *MSH2*, *PMS2* and *MSH6*, are typically implicated in the transmission of Lynch syndrome.

Large scale germline deletions within the mismatch repair gene *MSH2* account for a significant proportion of all deleterious mutations in this gene, and cause HNPCC. A recently characterized deletion of exons 1 through 6 of *MSH2* has been associated with a founder event within the United States which genealogical studies have previously dated to 1727.[2] The sequencing of the breakpoints completed the characterization of the "American Founder Mutation" (AFM) and allowed for the separation of this deletion from that of other similar, but distinct, deletions via long-range PCR and sequencing.[3,4] Here, we report the development of a robust multiplex PCR which has assisted in the detection of 24 new families who carry the AFM and the subsequent characterization of 20 patient haplotypes which flank the deletion. New genealogical data has identified a likely erroneous connection in the original pedigree analysis. As such, three further families have been linked to the original pedigree and we have now identified an alternative founding family dating to

1764, as well as five further "subfounder" groups, all of which date to the eighteenth century.

The use of molecular data, combined with extensive pedigree analysis, provides independent measures of the origins of the AFM, with significant implications for the mutation's modern prevalence in America, which has been estimated to be nearly 19,000 individuals.[5] In this study, I have sought to gain further insight into the origin, occurrence, and spread of the AFM through improved genealogical and molecular genetic studies.

****

Introduction:

All colorectal cancer (CRC) begins as a polyp which, if caught through screening methods such as colonoscopy, can be removed before becoming cancerous. As such, CRC is a truly preventable form of cancer, yet it is among the four most common forms of cancer in the United States, with approximately 150,000 individuals diagnosed each year.[1] This makes CRC screening an important component of medical care from both a patient health viewpoint and an economic viewpoint.

Hereditary nonpolyposis colorectal cancer (HNPCC) is thought to account for upward up 3% of all CRC in the western world, and is the most common form of heritable CRC.[1] It is caused by germline mutations in the mismatch repair (MMR) genes MLH1, MSH2, PMS2 and MSH6. As MMR genes are responsible for repairing errors that occur naturally during DNA replication, loss of function mutations in these genes lead to an accumulation of mutations in the rest of the cellular genome.[6] When these mutations occur in other important genes which control transcription or apoptosis, the

cell may divide aberrantly forming an adenoma, and can eventually progress to form a HNPCC-associated cancer, including: colorectal, endometrial, renal pelvic, ovarian, stomach, and small bowel cancers, among other less common presentations.[7]

Mutations in the MMR gene *MSH2* confer an autosomal dominant cancer predisposition with an ~80% lifetime risk of HNPCC-associated cancers,[4] thus it is of the utmost importance to identify these at-risk patients so as to provide them with specifically focused medical examinations and treatments. As *Alu* repeats, which are highly repetitive sequences found in large numbers (100-500,000) in the human genome, occur frequently in *MSH2*,[8] the gene is highly susceptible to large genomic deletions which result from non-homologous recombination. In America, the most common of these genomic rearrangements is a deletion of exons 1-6, which Baudhuin et al. (2005) found in 22 of 86 (25.6%) *MSH2* deletion-positive samples.[9] 1-6 deletions have also been identified in many other countries including Poland, England, Italy, Holland, and Australia.

In the United States a recurring exon 1-6 deletion of *MSH2* has been identified. This Alu-repeat mediated deletion was originally identified in 7 apparently unrelated American families with Lynch syndrome at the Creighton University School of Medicine.[2] This discovery led to the screening of another 11 families, these at The Ohio State University, of which 2 possessed the same gene deletion.[2] The existence of an identical large deletion such as this suggested that a founder mutation may exist in the United States, wherein an original progenitor introduced a novel mutation into the gene pool, and because of the small, rapidly growing American population, the aberrant-gene frequency became over-expressed.

Using genealogical data from the families possessing the AFM, 3 of the 9 families were purportedly linked through 9-10 generations to a common founder couple, who immigrated to America from Hesse, Germany in the early 1700's.[2]

The ultimate aim of my project has been to discover the age of the AFM in America, in order to accurately determine the modern disease prevalence which has been estimated at nearly 19,000 individuals[5]. Through the use of genealogical research, we have attempted to identify more accurately the true progenitor of the mutation by finding common ancestors who are shared by multiple AFM probands and to reconcile these ancestors with those proposed by Lynch et al., 2004. Additionally, we have utilized microsatellite markers to create patient haplotypes for twenty AFM probands. This molecular data can then be used to statistically determine the age of the mutation, based on the size of the conserved region and known rates of recombination. Finally, we have worked to develop a more robust PCR diagnostic tool to utilize in identifying additional probands who carry the AFM. An account on this work and my findings constitutes the remainder of this report.

****

Methods:

PCR Based Screening and Breakpoint Characterization: Previously published primer sets (Nakagawa *et al*., 2003; Wagner *et al*., 2003) for this procedure were not ideal due to their tendency to produce false negative results primarily because the product was relatively large and spanned across three complete and one partial *Alu* repeat. For this study, we positioned the forward primer (5′-GCCTGGCGTCAAACGTT-3′) in a region between two repeat elements that lacked homology to other human sequences.

The reverse primer (5′-TGAGTCATTTTGGGGATCAGTT-3′) is very similar to the F3 primer utilized by Wagner et al (2003), however in combination with this new forward primer it produces a breakpoint-specific 562bp amplicon under standard PCR conditions. In order to ensure that poor DNA quality was not a reason for a negative result, this amplicon was multiplexed with an 811bp amplicon from the *BRAF* locus (For: 5′-AAGCATCTCACCTCATCCTAACACA-3′, Rev: 5′-GGATCACACCTGCCTTAAA TTGCAT-3′), in a 15µl PCR reaction. Each 15µl reaction, containing 7.5µl of GoTaq® master mix (Promega), 25ng of genomic DNA and 5 pmoles of each of the 4 primers, was cycled under the following conditions: 95°C for 2 min., 30 cycles of 95°C for 30 sec., 60°C for 30 sec. and 72°C for 45 sec., and a final extension at 72°C for 8 min. Samples with a breakpoint specific product were treated with ExoSAP-IT (USB Corporation) and sequenced with the forward primer to confirm the presence of an AFM specific breakpoint. (see *Figure 1* below)

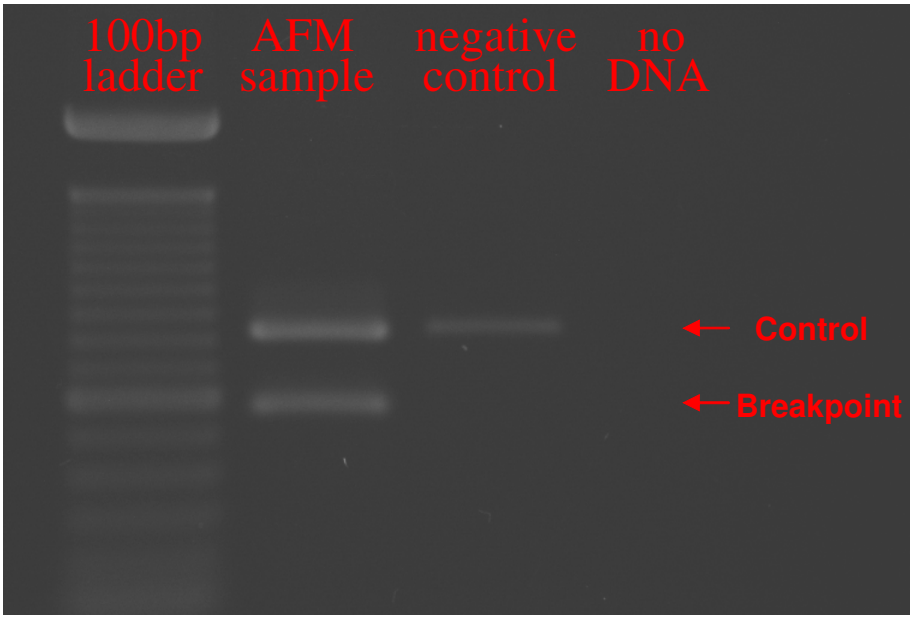


Figure 1: Agarose gel showing breakpoint specific PCR product

Genotype Analysis: A combination of 8 novel and 4 DeCode microsatellite markers were used to obtain a haplotype spanning ~12.48Mb across the *MSH2* locus. Microsatelite markers were chosen in lieu of SNPs because the increased number of possible alleles allowed for more precise differentiation between genetic conservation due to chance and that due to shared descent. Markers (see *figure 2*) were typed in 20 AFM positive probands (or their AFM positive kindred) and a set of 118 Coriell control DNAs.

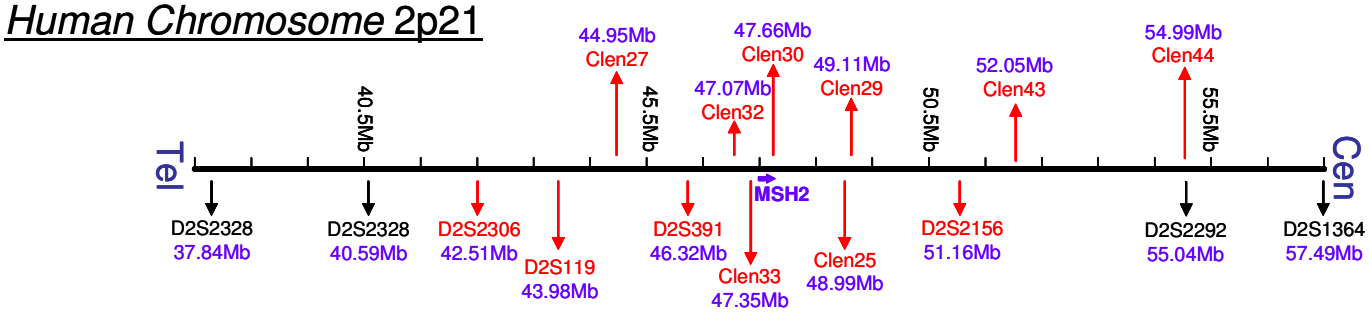


Figure 2: Graphical representation of the physical location of microsatellite markers flanking the *MSH2* gene. Markers in red were utilized in the haplotype analysis. Novel and database markers are prefixed by Clen and D2S, respectively.

For each amplicon the reverse primer had an M13 tail which, in combination with a FAM labeled M13 oligo, could be sized using an ABI7000. Each 25µl PCR reaction contained 12.5µl of AmpliTaq Gold PCR master mix (Applied Biosystems), 25ng of genomic DNA, 10 pmoles of forward primer, 2 pmoles of tailed reverse primer and 10 pmoles of the FAM-labeled M13 primer. Reactions were multiplexed when possible and cycled using the following profile: 96^{o}C for 10 min., 50 cycles of 96^{o}C for 30 sec., 60^{o}C for 30 sec. and 72^{o}C for 30 sec., and a final extension at 72^{o}C for 10 min.

The amplified product was sent for genotyping to determine the length of the labeled DNA amplicon. Polymorphisms in amplicon size, corresponding to differing numbers of microsatellite repeats, were scored as distinct alleles. The AFM-associated haplotype was inferred from a single haploid clone which was produced in a previous study by conversion to haploidy through fusion to a mouse recipient cell line.[10]

Genealogical investigations: Historical records were utilized to construct pedigrees for all probands, focusing on those lines marked by heightened HNPCC-associated cancer frequencies. The pedigrees were analyzed to search for any relatives shared by multiple probands, and also for any relatives that descended from the purported founder couple. Data were obtained from state and county birth, death, and marriage records, and from online genealogy resources such as those provided by Ancestry.com and FamilySearch.org. Genealogical data were also obtained through patient interviews conducted at a genetic counseling center, or with patient permission, through a telephone interview.

The lineage of the purported founders was also traced in order to link their descendants to the AFM families. Because this couple was instrumental in the creation of a German-Reformed Church in NC, the information concerning their descendants came from church records and local North Carolina genealogy enthusiasts, as well as the previously published research.[2] In addition, a research team traveled to this church during an annual family reunion held there to gather ancestry information and perform genetic testing on 45 descendants of the purported founder.

****

Results:

Identification of AFM carriers: The newly developed PCR was very successful in identifying AFM-positive individuals. The new primers resulted in no false negatives, unlike the previously published primers, and show great promise as a simple and inexpensive diagnostic tool for AFM screening. With the aid of this diagnostic, we have identified 24 new families with the AFM. Combined with the nine from the original study,[2] we have identified 33 families in all.

Due to the sequence composition of the genomic region containing *MSH2* (i.e. excessive numbers of *Alu*-like repeats), it is essential to confirm the exact breakpoint sequence of this deletion as it is very likely that other 1 through 6 deletions may occur as a result of recombination between alternate repeat elements or even at different locations within the same repeat elements.[11] Each of these 33 probands have tested positive for the AFM using the new primers, save six of the seven original probands whose mutation status was confirmed using the previously published primers.

For each of the twenty probands whose DNA was used in the haplotype analysis, the samples were amplified using the new breakpoint specific primers and the products were sequenced to confirm the precise location of the break. All 20 of the samples were found to carry the same AFM specific mutation (see *figure 3*).

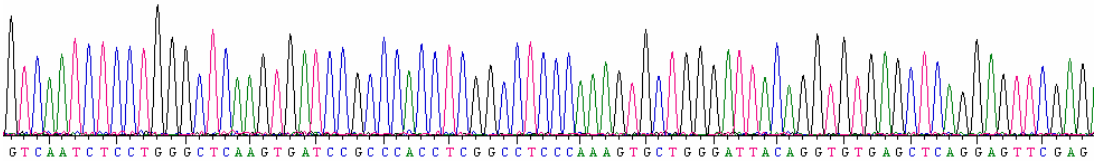


Figure 3: Breakpoint sequencing chromatogram. Alignments above represent wild type sequences 5′ and 3′ to the breakpoint, with mismatches shown in a red font. The expected breakpoint region (10bp) is depicted in a bold font.

Genealogical Studies: To support the genetic evidence it was hoped it would be possible to link some or all of the proband families to the 18th-century German founders hypothesized by Lynch et al (2004). Careful analysis showed, however, that one of the three genealogical links made in this earlier study was based on inaccurate genealogical data, causing the likely misidentification of a founder. However, the other two families are, in fact, genealogically linked. Furthermore, we have now shown how three additional families from our ongoing study connect to those from the initial study (see *figure 4*).

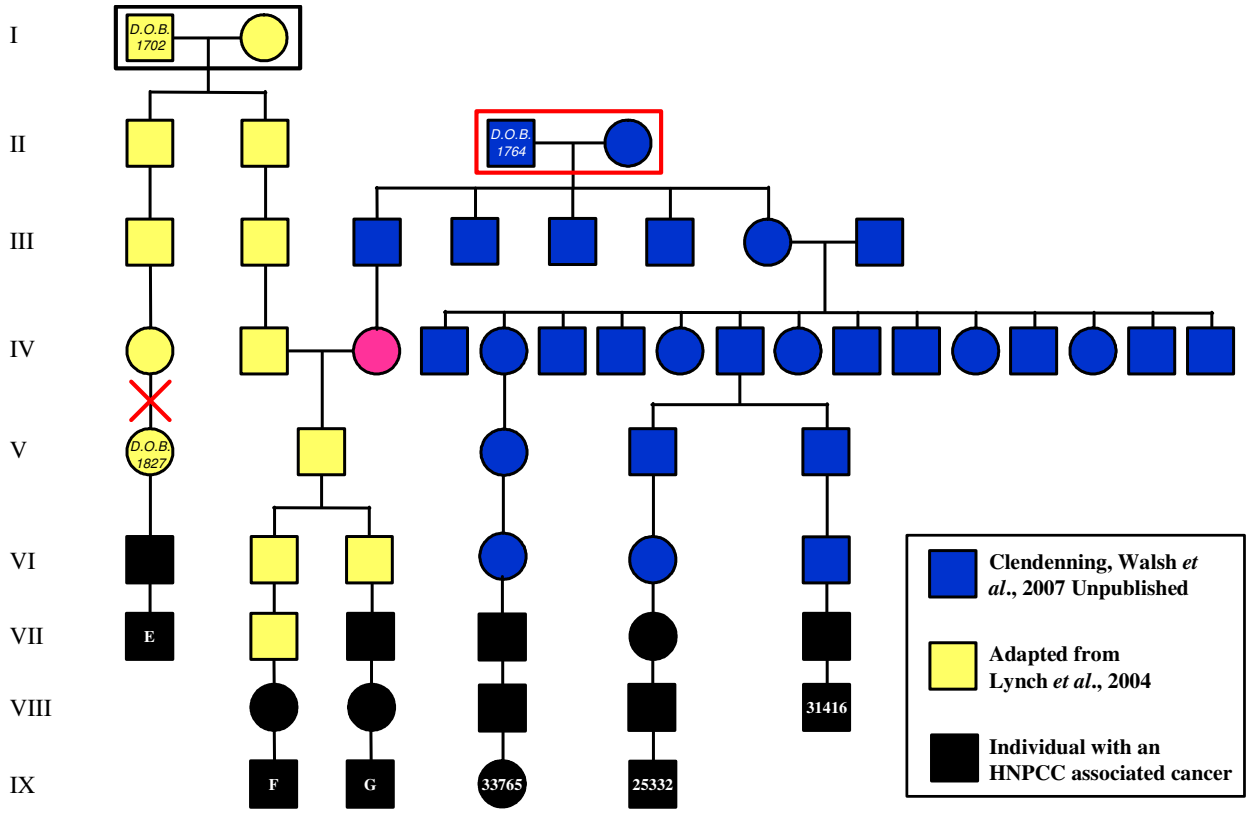
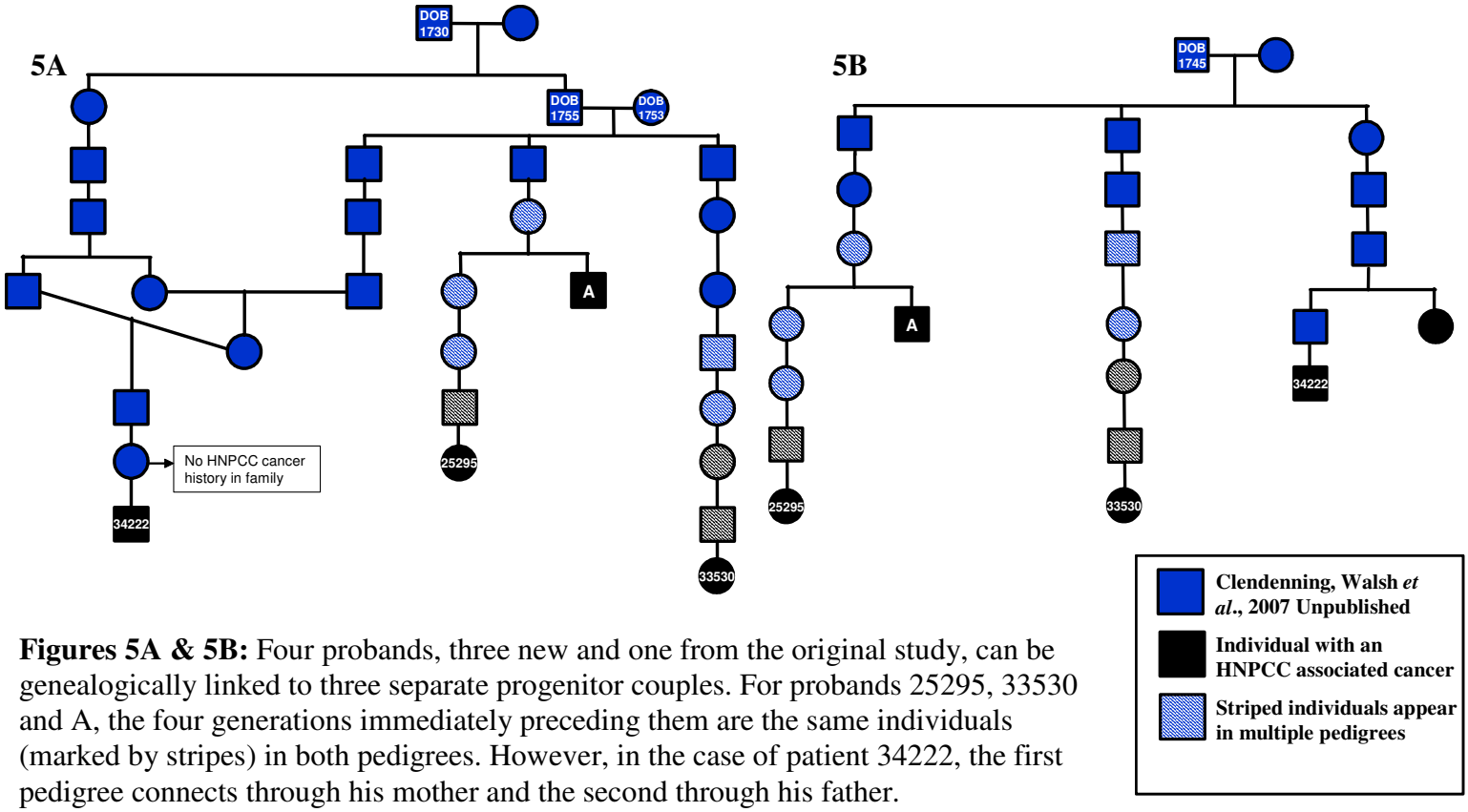


Figure 4: Pedigree showing the newly characterized genealogical link between three new probands and two of the originally linked families (Lynch *et al.*, 2004). The female shown in pink depicts the common link between the two original families (F and G) and the three newly identified families (25332, 31416, and 33765). Evidence obtained in our studies has identified a likely erroneous link in the original pedigree (marked by a red X), which would suggest that the founder couple proposed by Lynch *et al., 2004* (boxed in black) was misidentified. The new genealogical data presented here can trace the mutation back to one of two individuals (boxed in red). Currently, no data has been obtained for the female; however, we were able to trace all 4 of the male progenitor's grandparents back to central Europe (Germany and Switzerland).

Of the 42 descendants of the previously proposed founders, each of whom was tested for the AFM during a family reunion held in North Carolina during the summer of 2005, none were found to carry the mutation. This provides further evidence that the proposed German progenitors are part of a separate family tree which is only linked to the true AFM line by a single marriage event which occurred in 1832 (see *figure 4*)

In addition to the above pedigree, 11 more families have been linked to each other through a total of six additional ancestral couples. I shall refer to these other pedigrees as "subfounder" lines, all of which have an oldest known common ancestor born in the United States during the 1700's. While some of these couples actually predate those depicted in *figure 4*, I will continue to treat *figure 4* as the primary pedigree because it connects the largest number of probands (5) to a single ancestral couple.

The most interesting of the subfounder pedigrees connects 4 probands, 3 new and 1 from the original study, to three separate ancestral couples dating to 1730, 1755 and 1745 respectively (see *figure 5*).



Figures 5A & 5B: Four probands, three new and one from the original study, can be genealogically linked to three separate progenitor couples. For probands 25295, 33530 and A, the four generations immediately preceding them are the same individuals (marked by stripes) in both pedigrees. However, in the case of patient 34222, the first pedigree connects through his mother and the second through his father.

Three plausible progenitor couples exist in the pedigree of these families as the result of a highly intermarried, extremely sedentary family history. As seen in *figure 5*, the first several generations preceding probands 25295, 335530, and A are the same individuals in both pedigrees. Proband 34222, however, is connected through his mother in the first pedigree (*5A*) and through his father in the second (*5B*). Although all three couples could be the one who passed along the AFM, there is evidence to point to the couple dating to 1745, depicted in the second pedigree. This is because patient 34222 had no HNPCC-related cancer history on his mother's side of the family. While his father did not have an HNPCC-associated cancer either, his paternal aunt had colon cancer. This family history supports the hypothesis that it is the couple dating to 1745 (*5B*) who in fact passed on the AFM.

Figures 6 through 8 depict three more pedigrees, which link an additional 7 families (5 new, 2 from the previous study) to common ancestral couples dating to 1758, some time before 1800, and 1776 respectively (see *figures 6,7,* and *8*).

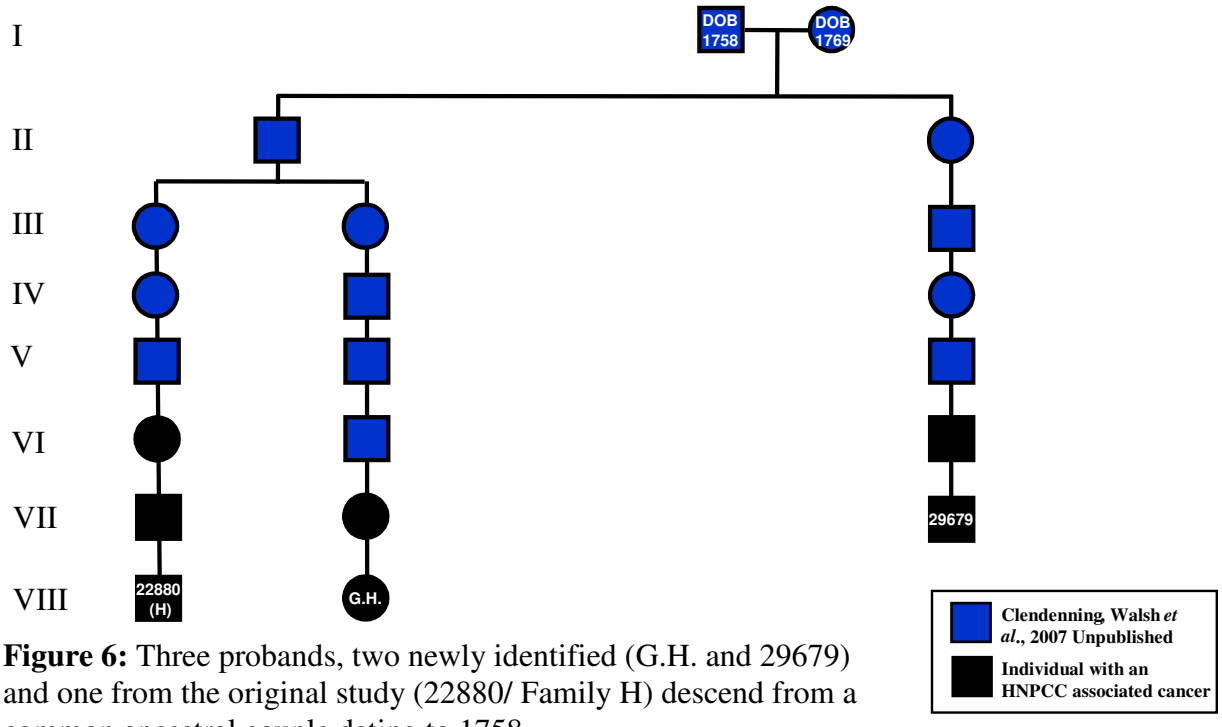


Figure 6: Three probands, two newly identified (G.H. and 29679) and one from the original study (22880/ Family H) descend from a common ancestral couple dating to 1758.

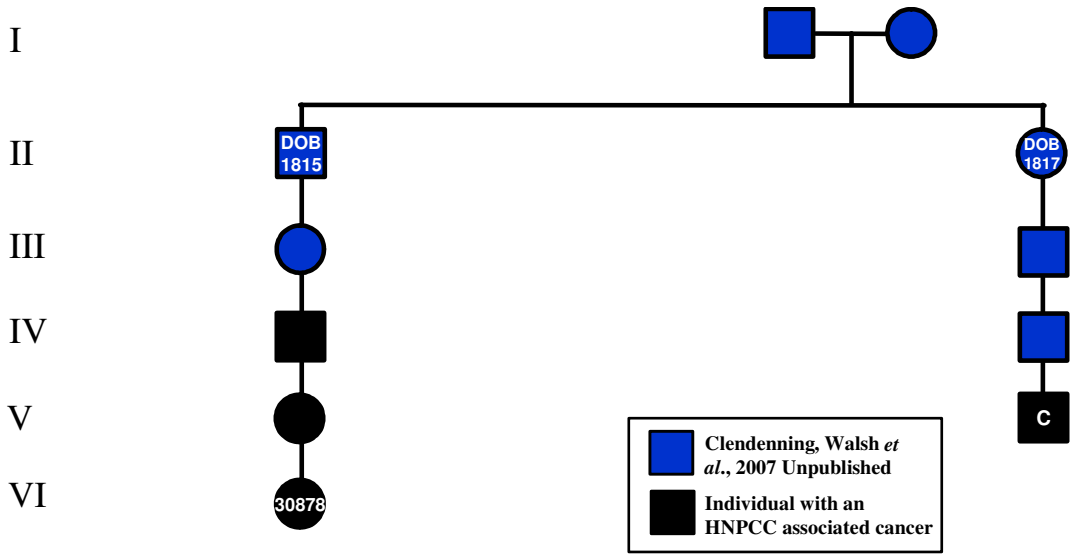


Figure 7: Two probands (one newly identified (30878) and one from the original study (C)) descend from a common ancestral couple predating the 18th century.

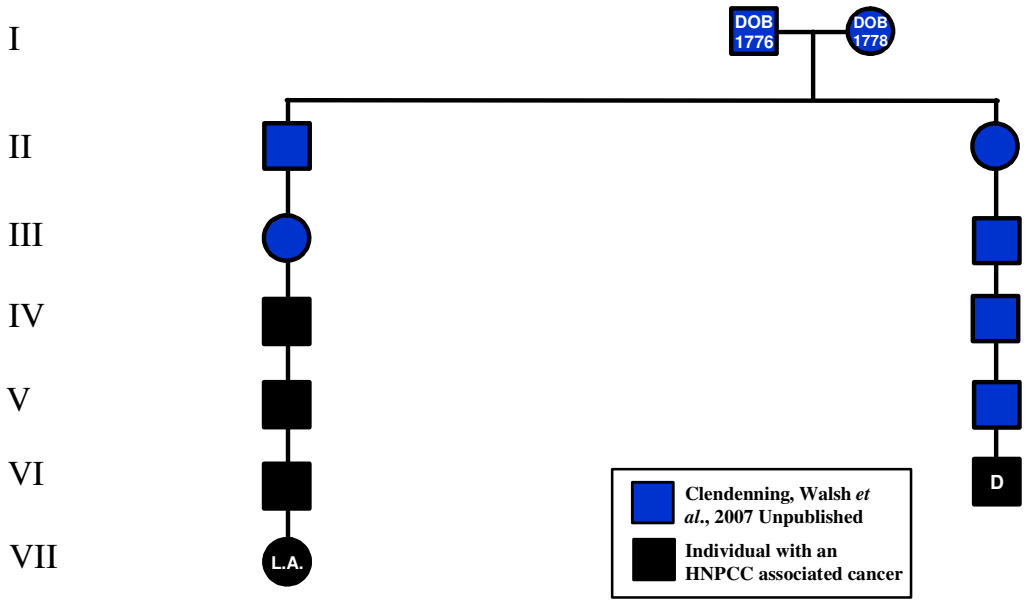


Figure 7: Two probands (one newly identified (L.A.) and one from the original study (D)) descend from a common ancestral couple dating to 1776.

Despite the fairly long existence of the AFM in America, which dates back to at least 1755 based on pedigree analysis (see *figure 5A*), it would be presumptuous to assume that the mutation has dispersed equally across the U.S. As is the case with other

founder mutations[12, 13, 14], the AFM appears to be over-expressed in certain regions. By looking at a proband's family history, one can typically determine which side of the family carries the AFM. Plotting the ancestral residence of this side of the family tree on a map of the U.S. shows us that certain states contain a much higher proportion of AFM ancestral kindred than others (see *figure 8*). The distribution of the AFM that is seen likely has historical explanations.

One example of this is that three Texan families and a Californian family are all descended from a common female ancestor (see *figure 4*, marked in pink). This individual married a Mormon convert who moved from Tennessee to Alabama, before assuming a leadership role in the Church of Latter Day Saints in Utah. Following an excommunication from the church, he fled to California where he resided until his death. As shown from this example, a single individual who carried the AFM early in America's history left a host of mutation-positive ancestors across the western United States.

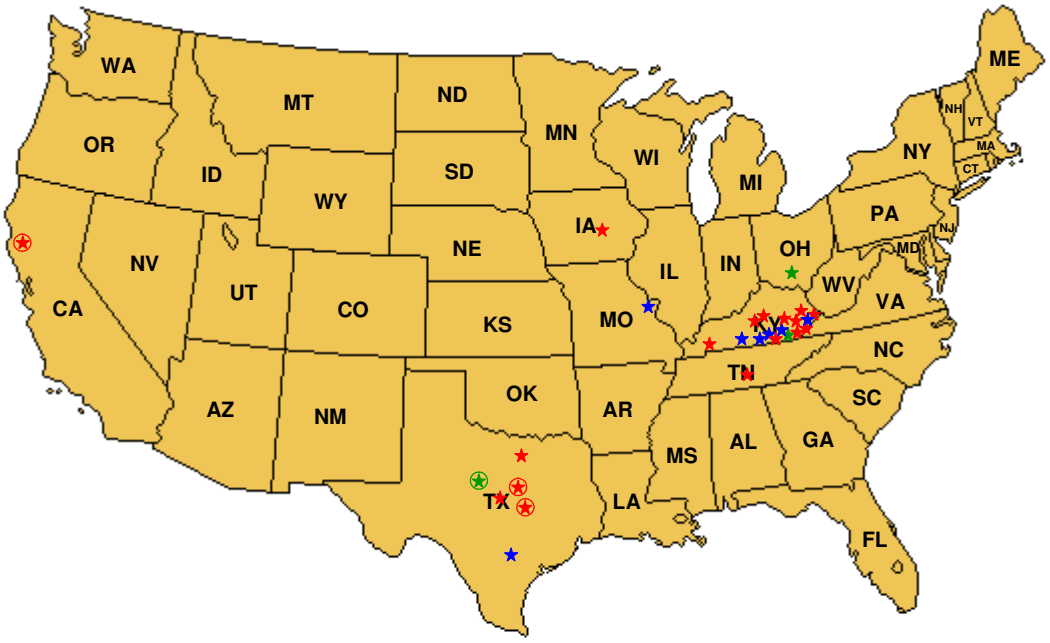


Figure 8: Map of the United States showing, where possible, the location of the proband's ancestral origins (used when the mutation-positive side of the family is clear or all four grandparents are from the same region). Red and green stars represent families that have been characterized in this study. Families that were also typed in the original study are shown in green. The great grandson of individual E is depicted as a green circled star. Three probands: 25332, 31416, and 33765 (represented by red circled stars), were genealogically linked to families F and G from the original study (see *figure 4*).

As 16 probands trace their ancestral origins to Kentucky, it is abundantly clear that the AFM is overexpressed in that state, especially the Southeastern region. For a more detailed view of the state, see *figure 9*. Bell County, KY in particular has an abundance of historical individuals who carried the AFM, as five probands trace their ancestry to that one county. This happens to be the county where the Cumberland Gap is located, which was the main passage for pioneers heading west from Virginia and the Carolinas. This supports our hypothesis that the AFM is a founder mutation which came to America from Europe and later spread westward with early settlers and is not, instead, a much older aboriginal mutation.

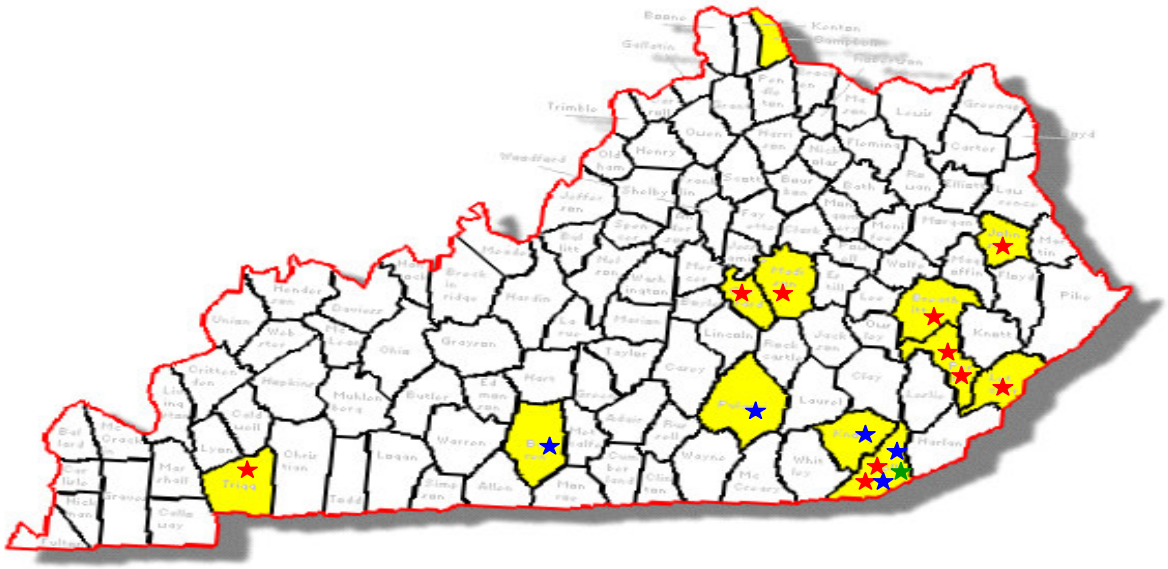


Figure 9: Expanded map of Kentucky showing, by county, the location of 16 proband's ancestral origins (used when the mutation-positive side of the family is clear or all four grandparents are from the same region). Red and green stars represent families that have been characterized in this study. Families that were also typed in the original study are shown in green. Highlighted counties include Barren, Bell, Campbell, Breathitt, Garrard, Johnson, Knox, Letcher, Madison, Perry, Pulaski, and Trigg.

Genotype Characterization: To link the probands genetically, and rule out the possibility that affected families were acquiring the mutation *de novo*, we typed a set of

polymorphic microsatellite markers which flanked the mutation site. A panel of 12 microsatellites was chosen which flanked the mutation site and had an average interval of 956Kb (range: 120Kb-2Mb). We optimized microsatellite selection based on three criteria: 1) the repeat had to be highly polymorphic to reduce the likelihood of similarity by chance 2) the microsatellite and its flanking sequence had to be unique within the genome 3) the repeat could not be 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 the excessive prevalence of *Alus*.

The use of a haploid cell line from one of the patients, 20600, enabled us to generate an inferred haplotype for all 20 patients (see *figure 10*) and demonstrated that all patients shared a common disease haplotype. Analysis demonstrated a conserved haplotype of some 2.67-7.10 Mb in most cases, but a much smaller haplotype (0.59-2.67 Mb) in a single case, which would formally indicate that the mutation is considerably older than initially believed.

Because we have hypothesized that patients possessing the AFM inherited it from a common founder, we expect the haplotype data to show that the 20 individuals are more closely related than a random outbred sample of Americans. Additionally, we anticipated that our within-group variance would be lower between individuals who have been genealogically linked to one-another, as they would ostensibly be more related to each other than to any of the other AFM patients. Looking at the haplotypes of the subgroups that have been linked, we can see that they do, in fact, most resemble the haplotypes of the other patients to which they have been genealogically linked (see *figure 11*).

Patient ID	D2S2306		D2S119		Clen27		D2S391		Clen32		Clen33		AFM	Clen30		Clen25		Clen29		D2S2156		Clen43		Clen44	
20600†	4		7		1		2		4		1			2		6		2		4		3		4	
20600	4	1	7	7	1	6	2	2	4	7	1	5	Yes	2	2	6	10	2	6	4	3	3	6	4	7
25480	4	1	6	7	1	1	2	5	4	3	1	1	Yes	2	2	6	4	2	7	4	4	3	5	1	2
26811	4	1	6	4	1	6	2	2	4	6	1	2	Yes	2	2	6	1	2	5	4	5	3	4	4	4
22880	1	2	6	3	1	1	2	5	4	3	1	3	Yes	2	1	6	8	2	5	4	6	3	5	4	7
25208	1	1	6	4	1	1	2	2	4	3	1	3	Yes	2	4	6	9	2	5	4	2	3	3	4	2
25443	1	1	6	4	1	1	2	1	4	7	1	2	Yes	2	3	6	10	2	5	4	5	3	5	2	3
25954	1	3	6	3	1	4	2	3	4	1	1	3	Yes	2	3	6	9	2	5	4	4	3	5	2	3
29679	1	1	4	2	1	3	2	1	4	4	1	1	Yes	2	3	6	7	2	5	4	4	3	6	4	8
33859	1	1	4	5	1	2	2	2	4	4	1	1	Yes	2	1	6	2	2	5	4	4	3	5	2	2
30462	4	3	4	4	1	6	2	2	4	7	1	1	Yes	2	3	6	10	2	5	4	5	3	4	2	6
26737	1	1	6	1	1	1	2	4	4	4	1	1	Yes	2	1	6	11	2	5	4	5	5	6	2	7
25332	1	2	6	5	1	7	2	4	4	4	1	5	Yes	2	1	6	3	2	6	2	5	3	5	2	5
31416	1	1	4	5	4	6	2	3	4	4	1	1	Yes	2	1	6	8	2	4	2	5	3	5	5	5
24342	4	4	1	5	2	6	2	4	4	5	1	3	Yes	2	3	6	8	3	3	2	5	3	5	5	5
33765	1	3	5	5	1	6	2	2	4	4	1	3	Yes	2	1	6	3	1	6	2	7	1	5	2	6
34222	1	1	*	*	1	6	2	2	4	3	1	1	Yes	2	1	5	9	5	5	5	6	3	5	4	6
29649	1	1	1	4	1	1	4	5	4	4	1	1	Yes	2	2	7	9	5	5	1	6	3	5	4	2
33530	1	2	1	4	1	1	2	5	4	2	1	4	Yes	2	3	7	10	1	5	3	6	5	6	4	5
25295	1	1	6	1	1	1	2	2	4	4	1	3	Yes	2	1	5	9	5	6	6	5	5	5	4	6
30878	1	4	3	3	1	3	2	2	4	7	1	1	Yes	2	2	2	11	5	6	2	6	4	6	4	3
Population frequency of disease allele	n/a		n/a		37%		30%		47%		58%			26%		3%		3%		17%		12%		n/a	

Figure 10: Genotype data for 20 AFM individuals. Conserved disease haplotype is highlighted in yellow (data inferred from a single haploid clone†). Allele frequencies are determined from a set of 118 control DNAs.

Patient ID	D2S2306		D2S119		Clen27		D2S391		Clen32		Clen33		AFM	Clen30		Clen25		Clen29		D2S2156		Clen43		Clen44	
33765	1	3	5	5	1	6	2	2	4	4	1	3	Yes	2	1	6	3	1	6	2	7	1	5	2	6
25332	1	2	6	5	1	7	2	4	4	4	1	5	Yes	2	1	6	3	2	6	2	5	3	5	2	5
31416	1	1	4	5	4	6	2	3	4	4	1	1	Yes	2	1	6	8	2	4	2	5	3	5	5	5
34222	1	1	*	*	1	6	2	2	4	3	1	1	Yes	2	1	5	9	5	5	5	6	3	5	4	6
25295	1	1	6	1	1	1	2	2	4	4	1	3	Yes	2	1	5	9	5	6	6	5	5	5	4	6
33530	1	2	1	4	1	1	2	5	4	2	1	4	Yes	2	3	7	10	1	5	3	6	5	6	4	5
22880	1	2	6	3	1	1	2	5	4	3	1	3	Yes	2	1	6	8	2	5	4	6	3	5	4	7
29679	1	1	4	2	1	3	2	1	4	4	1	1	Yes	2	3	6	7	2	5	4	4	3	6	4	8
Population frequency of disease allele	n/a		n/a		37%		30%		47%		58%			26%		3%		3%		17%		12%		n/a	

Figure 11: Genotype data for 8 AFM individuals pulled from *figure 10*. These haplotypes are divided into three subgroups that have been demonstrated to share common ancestors. Patients 33765, 25332 and 31416 share a common ancestral couple 6 generations removed (see figure 4). Patients 34222, 25295 and 33530 share 3 common ancestral couples who are 5-7, 6-8 and 7-9 generations removed (see figures 5A and 5B). Patients 22880 and 29679 share a common ancestral couple 6-7 generations removed (see figure 6).

****

Discussion:

In this study we identified 24 additional families that carry the AFM. Coupled with the nine families from the previous study this makes a total of 33 families spread throughout the United States. As mutations in *MSH2* account for a significant proportion (1-2%) of all cases of colorectal cancer[1], it is important to assess the risk that can be attributed to specific frequent mutations. Additionally, genomic rearrangements constitute a large proportion (10-20%)[3] of all mutations in *MSH2*, and have been reported to constitute even larger proportions of some samples (45.4-50.0%) [4,9]. The AFM appears to be the most common of these rearrangement mutations (33-75%).[9,4]

By using both genealogical analysis and haplotype data, we hoped to independently determine both an historical and a molecular estimate for the age of the mutation. At this time we can conclude, based on the genealogy, that the AFM has existed in the United States for at least 250 years (see *figure 5A*). Five pedigrees have identified a total of seven couples, each of which is ancestral to multiple AFM patients. These ancestors lived sometime before 1800 and are likely connected by pedigree to a still earlier couple who is ancestral to all our families.

Genotype analysis of 20 individuals demonstrated a conserved haplotype of some 2.67-7.10Mb in most cases, but a much smaller haplotype (0.59-2.67Mb) in a single case, which would indicate that the mutation is, perhaps, considerably older than initially believed. Given that the AFM has not been seen outside the United States, we are left with two possible explanations for our findings.

1.) The mutation is older than previously thought and exists outside the United States in regions in which it has not yet been detected. In this case the mutation may have been introduced into the U.S. by two or more individuals.

2.) The short haplotype (29649) is a result of frequent recombination events occurring within and around *MSH2* due to the many *Alu* repeat elements, or due to microsatellite mutations, which occur in dinucleotide repeat elements at a frequency in the range of 10^{-3}-10^{-4} times per locus per meiosis.[15] Given that the haplotype data contains a total of 12 markers genotyped on 20 patients who descend from a shared ancestor at least 10 generations removed from our probands, this provides a minimum 200 meiotic divisions and thus 2400 opportunities for a single germline MSR mutation to have occurred (20x10x12). Thus, it is likely that some of our patient markers differ from the disease haplotype due to MSR mutations as well as *Alu*-mediated recombination events. These phenomena may explain the short haplotype, in which case the AFM could have been introduced by a single founder who was de novo for the mutation or whose blood relatives did not pass the AFM on to descendants in Europe where we have screened for the AFM.

The age of the mutation based upon the molecular data remains to be determined, but considering what is now known, prevalence estimates based on a single European founding event predating the 1750's should provide a conservative but reasonable estimate of the true prevalence of the AFM.

The spread of the AFM can be qualitatively viewed in several ways. Based on ancestral locations, we see it appearing in 7 states. Today it can be found in 17 states, which contain more than half of the American population, as well as the U.K. (16 states

identified by Lynch et al (2006), plus newly identified patients in Iowa and England). Overall, the spread of the AFM seems substantial and only stands to increase over time.

With a founding event predating 1750 in a nation which underwent rapid population growth and westward expansion due to large family sizes and increased mobility, the AFM has become a widespread mutation. It is perhaps the second most common mutation in *MSH2*, after the intronic germline A-->T point mutation at nt942+3, which frequently occurs de novo and accounts for 11% of all known pathogenic MSH2 mutations.[16]

Because it is extremely important to screen for CRC in individuals with MMR mutations, it is vital that these mutations be identified and that the frequency of colonoscopy screenings for such individuals be increased to occur at least biannually starting at age 20, as in accordance with the recommendations of the International Collaborative Group on HNPCC.[7] It is also vital to determine the nature of these MMR mutations so that family members can receive targeted genetic testing followed by increased screenings for those determined to have a germline MMR mutation.

In the case of the AFM, the newly developed diagnostic PCR provides an inexpensive and highly accurate method of detecting the mutation. As the AFM appears to be very common throughout the United States, especially in the Midwest and Texas, we propose that the new PCR test should be used as a method of primary mutation testing for individuals already found to have a loss of *MSH2* protein expression via immunohistocompatibility. Additionally, the test has applicability for patients with an undiagnosed *MSH2* mutation which was not elucidated by genomic sequencing, or for individuals determined to have a genomic deletion of *MSH2* as determined by Multiplex

Ligation-dependent Probe Amplification (MLPA[17]). This could prove especially effective in the regions determined to be geographic hotspots for the mutation and has real implications for cancer prevention and control.

References:

1. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med. 2003; 348(10): 919-32.
2. Lynch HT, Coronel S, Hampel H, de la Chapelle A, et al. A founder mutation of the MSH2 gene and hereditary nonpolyposis colorectal cancer in the United States. JAMA. 2004; 291: 718-724.
3. Nakagawa et al. Identification and characterization of genomic rearrangements of MSH2 and MLH1 in Lynch syndrome (HNPCC) by novel techniques. Hum Mut. 2003; 22(3): 258.
4. Wagner A, Barrows A, et al. Molecular analysis of hereditary nonpolyposis colorectal cancer in the united states: high mutation detection rate among clinically selected families and characterization of an American founder deletion of the MSH2 gene. Am J Hum Genet. 2003; 72: 1088-1100.
5. Lynch et al. American founder mutation for Lynch syndrome: prevalence estimates and implications. Cancer. 2006; 106 (2): 448-462.
6. Peltoma¨ki, P. Role of DNA mismatch repair defects in the pathogenesis of human cancer. Journ Clin Onc. 2003; 21(6): 1174-1179.
7. Möslein et al. Clinical aspects of HNPCC. Annals NY Acad Sciences. 2001; 910 (1): 75–84.
8. Charbonnier et al. The 5' region of the MSH2 gene involved in hereditary non-polyposis colorectal cancer contains a high density of recombinogenic sequences. Hum Mutat. 2005; 26(3): 255-61.
9. Baudhuin, et al. Characterization of hMLH1 and hMSH2 gene dosage alterations in Lynch syndrome patients. Gastroenterology. 2005; 129: 846-854.
10. Yan H, Papadopolous N, et al. Conversion of Diploidy to Haploidy: Individuals Susceptible to Multigene Disorders May Now Be Spotted More Easily. Nature. 2000; 403: 723-724.
11. van der Klift et al. Molecular characterization of the spectrum of genomic deletions in the mismatch repair genes MSH2, MLH1, MSH6, and PMS2 responsible for hereditary nonpolyposis colorectal cancer (HNPCC). Genes Chrom Cancer. 2005; 44(2):123-38.
12. Bar-Sade, et al. Could the 185delAG BRCA1 mutation be an ancient Jewish mutation? Eur J Hum Genet. 1997; 5(6):413-6.
13. Mykkanen, et al. Detection of the founder effect in Finnish CADASIL families. Eur J Hum Genet. 2004;12(10):813-9.
14. Founding mutations and Alu-mediated recombination in hereditary colon cancer. Nature Medicine. 1995; 1: 1203 – 1206.
15. Ellegren, H. Microsatellite mutations in the germline: implications for evolutionary inference. Trends in Genetics. 2000;16(12):551-8
16. Desai, DC, et al. Recurrent germline mutation in MSH2 arises frequently de novo. Journal Med Genetics. 2000; 37(9):646-52.

17. Schouten et al. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucelic Acids Research. 2002; 30(12): e57.