Am. J. Hum. Genet. 73:1157–1161, 2003

Report

Missense Mutations in *hMLH1* and *hMSH2* Are Associated with Exonic Splicing Enhancers

Ivan P. Gorlov, Olga Y. Gorlova, Marsha L. Frazier, and Christopher I. Amos

Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston

There is a critical need to understand why missense mutations are deleterious. The deleterious effects of missense mutations are commonly attributed to their impact on primary amino acid sequence and protein structure. However, several recent studies have shown that some missense mutations are deleterious because they disturb *cis*-acting splicing elements—so-called "exonic splicing enhancers" (ESEs). It is not clear whether the ESE-related deleterious effects of missense mutations are common. We have evaluated colocalization of pathogenic missense mutations (found in affected individuals) with high-score ESE motifs in the human mismatch-repair genes *hMSH2* and *hMLH1*. We found that pathogenic missense mutations in the *hMSH2* and *hMLH1* genes are located in ESE sites significantly more frequently than expected. Pathogenic missense mutations also tended to decrease ESE scores, thus leading to a higher propensity for splicing defects. In contrast, nonpathogenic missense mutations (polymorphisms found in unaffected individuals) and nonsense mutations are distributed randomly in relation to ESE sites. Comparison of the observed and expected frequencies of missense mutations in ESE sites shows that pathogenic effects of ⩾20% of mutations in *hMSH2* result from disruption of ESE sites and disturbed splicing. Similarly, pathogenic effects of ⩾16% of missense mutations in the *hMLH1* gene are ESE related. The colocalization of pathogenic missense mutations with ESE sites strongly suggests that their pathogenic effects are splicing related.

Missense mutations—nucleotide substitutions that change an amino acid in a protein—are among the most common types of mutations underlying inherited human diseases. The deleterious effects of missense mutations are usually attributed to their effects on protein function. However, recent studies of normal and alternative splicing suggest that the deleterious effects of nucleotide substitutions might, in fact, be splicing related when they are located in exonic splicing enhancers (ESEs) (Cartegni and Krainer 2002; Cartegni et al. 2002; Fackenthal et al. 2002; Moseley et al. 2002; Pollard et al. 2002). ESEs are discrete, degenerate motifs of 6–8 nts located inside exons (Liu et al. 1998; Blencowe 2000). The study of normal splicing suggests that most exons contain at least one functional ESE site (Blencowe 2000; Hastings and Krainer 2001; Cartegni et al. 2002). ESEs are target sequences for the family of conserved essential splicing factors—the serine- and arginine-rich (SR) proteins (Stojdl and Bell 1999; Graveley 2000; Hastings and Krainer 2001). ESEs play an important role in exon recognition. Nucleotide substitutions in ESEs can result in failure of SR proteins to bind to the ESE, which leads to failure of splisosome machinery to recognize the sequence as exonic and causes exon skipping (Ars et al. 2000; Cartegni et al. 2002; Fackenthal et al. 2002; Moseley et al. 2002). Each SR protein recognizes specific, albeit degenerate and partially redundant, sequence motifs. ESE motifs for four members of the SR family (SF2/ASF, SRp40, SRp55, and SC35) have been identified (Liu et al. 1998; Stojdl and Bell 1999; Graveley 2000; Liu et al. 2000). To identify the ESE motifs that are recognized by individual SR proteins, a PCR-based approach called "SELEX" (systematic evolution of ligands by exponential) enrichment was used. In this approach, a natural splicing enhancer in a minigene is replaced by short, random sequences derived from an oligonucleotide library. The generated pool of minigenes is transfected into cultured cells, and spliced mRNAs are amplified by RT-PCR and sequenced (Liu et al. 1998, 2000). On the basis of the frequencies of the individual

Received June 19, 2003; accepted for publication July 31, 2003; electronically published October 1, 2003.

Address for correspondence and reprints: Dr. Christopher I. Amos, Department of Epidemiology, Box 189, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030. E-mail: camos@mail.mdanderson.org

0002-9297/2003/7305-0017$15.00

nucleotides at each position, a score matrix for each nucleotide in each position was calculated. This score matrix can be used to predict SR protein–specific ESEs (ESEfinder).

We studied the association of pathogenic missense mutations (found in affected kindreds), nonpathogenic missense mutations (polymorphic mutations and sequence variants found in nonaffected individuals), nonsense mutations, and frameshifts (here restricted to 1- or 2-nt deletions and insertions) with ESE sites in *hMSH2* and *hMLH1,* human mismatch repair genes that are related to human nonpolyposis colon cancer (HNPCC [MIM 114500]) (Peltomaki and Vasen 1997). We used published and our own data on pathogenic and nonpathogenic missense mutations in the *hMSH2* and *hMLH1* genes (tables A, B, C, D, and E [online only]). Only mutations found in independent families were used. We excluded multiple reported mutations found in the same family. The numbers of different types of mutations analyzed are shown in table 1.

First, we searched the coding regions of the genes for the presence of ESE motifs with ESEfinder software. To reduce the number of false-positive results, we used a more-stringent-than-recommended threshold value of 3.0 for all four types of ESE motifs. Potential ESE motifs found in the *hMSH2* and *hMLH1* genes are listed in table E (online only). We excluded ESEs in exon/exon boundaries, accounted for overlap between different ESEs by counting as a single ESE any segment containing two or more ESEs, and estimated the percentage of sequence that consists of ESE motifs for each entire gene and each exon. This estimate provided us with the proportion of mutations expected to be in ESE motifs under the null hypothesis that assumes that there is no association between pathogenic missense mutations and exonic splicing enhancers.

Different nucleotide substitutions in mutation databases for the *hMSH2* and *hMLH1* genes differ with respect to how many times they are reported in the databases. Some of them are listed only once, whereas others are reported several times (e.g., the C→T transition at position 350 in the *hMLH1* gene is reported 11 times). Multiply reported mutations in the mutation databases originate from different families. Counting each reported mutation, we found that missense mutations are colocalized with ESEs (for *hMSH2,* $\chi^2 = 11.8$, df = 1, $P < .001$; for *hMLH1,* $\chi^2 = 7.9$, df = 1, $P < .01$) (fig. 1*a*). Alternatively, we also counted each separate mutation only once, no matter how often it was listed in the database. Again, we found that, in both genes, deleterious missense mutations are located in ESEs more frequently than expected (for *hMSH2,* $\chi^2 = 9.4$, df = 1, $P < .001$; for *hMLH1,* $\chi^2 = 4.3$, df = 1, $P < .05$). Counting each mutation only once eliminates all possibility that families are related but probably leads to a downward bias. Nucleotide substitutions can occur at any position in a coding region of a gene; yet only deleterious mutations that disturb important functional sites would lead to cancer and thus have a chance of being detected by screening of cancer-affected families. The more deleterious the mutation, the higher the chance it has of causing disease and of being detected. Therefore, deleterious mutations are expected to be most frequent in mutation databases (Martin et al. 2002; Olivier et al. 2002). Counting multiple mutations only once leads to loss of information and may cause downward bias by reducing the number of observations and by eliminating variation in the number of mutations between different mutant sites.

There may be two possible explanations for missense mutations in *hMSH2* and *hMLH1* genes being preferentially located in ESE motifs: either missense mutations arise more frequently in ESEs, or mutations in ESEs are more pathogenic than mutations outside ESEs and therefore are more likely to be detected during screening of affected individuals. Our analysis favors the second explanation. If a missense mutation becomes pathogenic because it is located in an ESE, then one can expect pathogenic missense mutations to be located in ESE sites more frequently, compared with nonpathogenic ones. We found that 57% (28/49) of pathogenic missense mutations were located in ESEs of the *hMSH2* gene versus 24% (4/17) of nonpathogenic mutations ($\chi^2_1 = 5.66$; $P = .02$; Fisher's exact test). For the *hMLH1* gene, we also found that pathogenic missense mutations were more frequently located in ESE sites than nonpathogenic mutations—58% (57/99) versus 38% (3/8), respectively ($\chi^2_1 = 1.21$; $P = .30$; Fisher's exact test)—but there were too few nonpathogenic mutations to draw a meaningful conclusion.

If deleterious effects of nucleotide substitutions located inside ESE motifs result from disruption of splicing, then

Table 1

Types and Numbers of Mutations Analyzed

	No. of Mutations			
Gene	Pathogenic Missense	Nonsense and Frameshift	Nonpathogenic Missense	Total Mutations
hMSH2	50	81	17	148
hMLH1	99	68	8	175

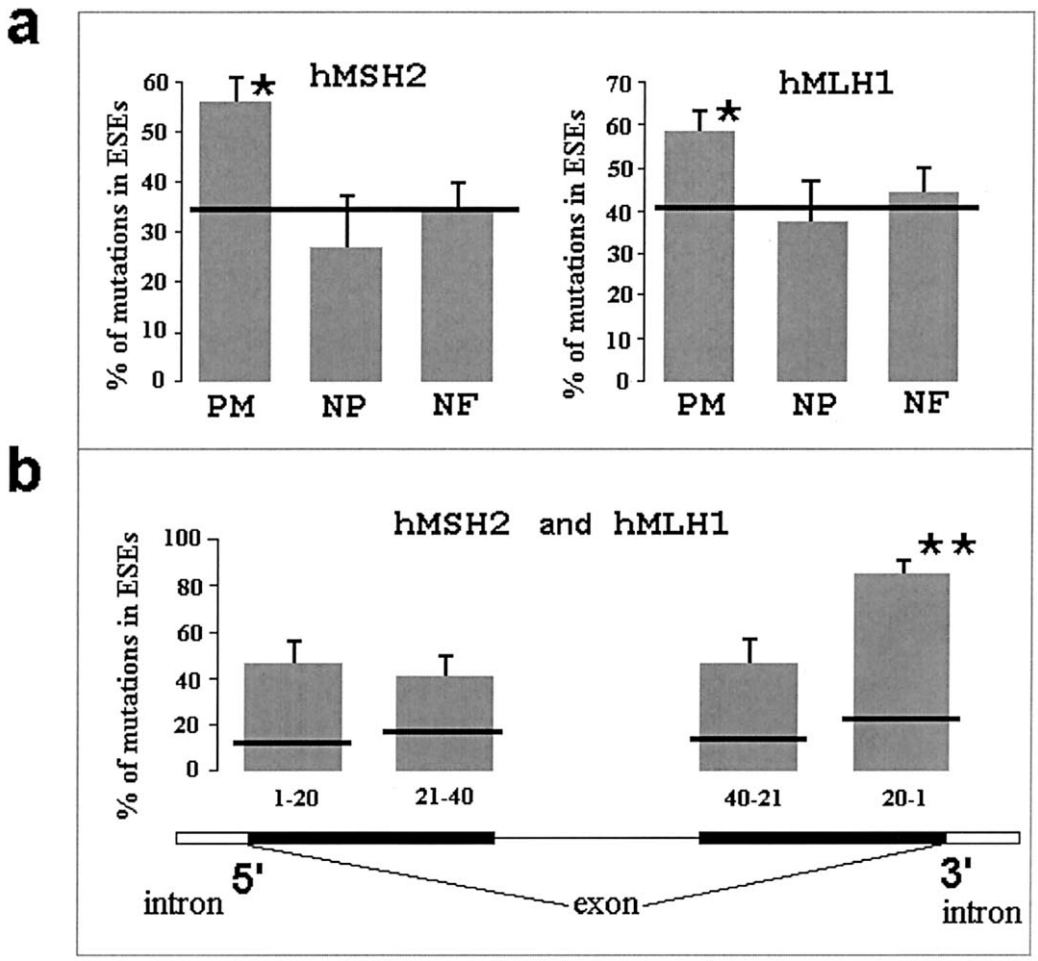

Figure 1 *a,* Frequency of ESE-associated mutations. Horizontal lines show the expected frequencies of mutations calculated as a proportion of the sequence occupied by ESE motifs. Differences between the expected and observed frequencies of pathogenic missense mutations (PM) were significant for both *hMSH2* and *hMLH1* (marked by an asterisk [*]). Nonpathogenic missense mutations (NP) show a trend to be underrepresented in ESEs. Nonsense and frameshift mutations (NF) did not show preferential association with ESE sites. Bars represent SEs of the frequencies. *b,* Positions of pathogenic missense mutations with respect to the 5′ and 3′ ends of an exon. The last 20 nts of exons had the highest frequency of mutations in ESEs: 83% of the mutations were in ESE sites (marked by a double asterisk [**]). The difference between the observed and expected numbers of mutations in that region was highly significant ($\chi^2 = 33.8$; df = 1; $P < .001$). When the analysis was limited to short (average size, 80 nts) exons, the colocalization of the pathogenic missense mutations with ESE sites was even higher: 96% of the mutations were in ESEs.

even missense mutations that are unlikely to affect protein structure (e.g., mutations that do not change the type of amino acid) will have a chance to be deleterious because they disturb ESEs and, therefore, splicing. This idea is supported by stratified analysis. We stratified missense mutations into "conservative" and "radical" (classifying them according to specifications of Dagan et al. [2002]), and we found that, in both genes, missense mutations located outside ESE sites tended to be "radical," strongly affecting protein-structure mutations, whereas those located in ESE motifs are more likely to be "conservative" mutations that have no or slight effect on protein structure. For *hMSH2* genes, the frequency of conservative missense mutations located in ESEs is 0.61 ± 0.09, whereas the frequency of conservative missense mutations outside ESEs is 0.50 ± 0.10. For the *hMLH1* gene, we found the same trend: the frequency of conservative missense mutations in ESEs is 0.31 ± 0.06, whereas the frequency of conservative mutations outside ESEs is 0.21 ± 0.10. For both genes, the differences in the proportions of conservative missense mutations located inside and outside of ESE sites are not significant, even after we combine the data for both genes, probably because of the relatively low number of mutations in analysis.

The correlation of different types of mutations with ESE sites can be explained as follows: nonsense and frameshift

mutations always produce truncated nonfunctional proteins and therefore always—no matter where they are located with respect to ESEs—are sufficiently damaging to cause disease. Thus, truncating mutations have a high chance of being detected by screening affected families. Missense mutations, especially those located outside important functional domains, may not change protein structure sufficiently to be pathogenic. However, if a nucleotide substitution occurs in a functional ESE site, it could disturb normal splicing and be sufficiently deleterious to cause disease. This could explain why affected individuals are enriched with missense mutations that are located in ESE sites and why polymorphisms found in unaffected individuals are not associated with ESEs and even show a trend not to be localized there.

Different single-nucleotide substitutions can change the ESE score in different directions: some substitutions increase the score, whereas others decrease it. If nucleotide substitutions located in ESE sites are deleterious because they disturb functional ESE sites, then such substitutions are expected mainly to decrease ESE scores. We compared the observed and expected proportion of score-decreasing missense mutations located inside ESE sites. The expected proportion of score-decreasing substitutions was calculated on the basis of all possible substitutions in the ESEs that lead to missense mutations. We found that ESE-located missense mutations reported in *hMSH2* and *hMLH1* mutation databases decrease ESE scores significantly more frequently than one would expect. For the *hMSH2* gene, we found that the expected frequency of score-decreasing mutations is 0.77 ± 0.03, whereas the observed frequency of score-decreasing mutations is 0.96 ± 0.04. The differences are highly significant ($\chi^2 = 6.5$; df = 1; $P < .01$). A similar result was obtained for the *hMLH1* gene: the expected frequency of score-decreasing mutations is 0.78 ± 0.02, whereas the observed frequency of score-decreasing mutations is 0.91 ± 0.06 ($\chi^2 = 5.9$; df = 1; $P < .01$).

The excess of pathogenic mutations in ESE sites compared with the expected frequency provides a minimal estimate of the proportion of missense mutations, the pathogenic effects of which are ESE related. As an upper limit for the estimate of the proportion of ESE-related mutations, one can suggest that all pathogenic missense mutations located in ESE sites are deleterious because they disturb functional splicing enhancers. This approach is likely to overestimate the proportion of ESE-related pathogenic mutations. First, not all ESE motifs are actual functional splicing enhancers (Cartegni 2002). Second, not all nucleotide substitutions in functional ESEs disturb their function (Cartegni and Krainer 2002; Fackenthal et al. 2002; Moseley et al. 2002; Pollard et al. 2002; ESEfinder). For the *hMSH2* gene, the observed frequency of pathogenic missense mutations in ESEs is 55%, whereas the expected frequency is 36%. This means that 20%–55% of missense mutations in *hMSH2* are pathogenic, because they affect ESE sites and therefore disturb normal splicing. A similar reasoning shows that the frequency of ESE-related mutations in the *hMLH1* gene is 16%–58%.

Although an exon usually has several ESE motifs, the splicing machinery does not use most of them (Cartegni 2002; ESEfinder). The question is whether the functional ESE sites are distributed randomly. If functional ESEs are preferentially located in some specific regions within an exon, then the association of the pathogenic missense mutations with ESEs will be higher in that region. A study of the molecular mechanisms of splicing suggests that functional ESE sites occupy specific positions relative to the 5′ or 3′ ends of an exon (Blencowe 2000; Hastings and Krainer 2001; Cartegni et al. 2002). We compared the expected and observed frequencies of pathogenic missense mutations in four regions: the first 20 nts located near the 5′ end of an exon, nts 21–40 near the 5′ end of an exon, the first 20 nts near the 3′ end of an exon, and nts 21–40, starting from the 3′ exon region (fig 1*b*). Because the number of missense mutations located in these specific regions is relatively low, we combined the data on both the *hMLH1* and *hMSH2* genes. We found that >80% of pathogenic missense mutations that are located in the last 20 nts of exons (especially in short exons—80 nts, on average) strongly colocalize with ESEs (fig. 1*b*). This finding suggests that functional ESEs are preferentially located near the 3′ ends of exons. However, since we used mostly short exons (~80 nts), it is noteworthy that, in fact, the functional ESEs are located 60–65 nts from the 5′ ends of exons.

Aside from the SR proteins that we have studied here, other classes of exonic enhancers exist; for example, those driven by hnRNP proteins (Chabot et al. 2003). In future research, studies of the frequency of mutations that affect other ESE sites would be of interest, once algorithms have been developed for evaluating the effect that mutations have on splicing related to these proteins. In summary, our analyses provide compelling evidence that many missense mutations associated with deleterious effects for *hMLH1* and *hMSH2* affect splicing. Of note, we found that conservative mutations that heretofore may not have had an obvious role in causing HNPCC may disrupt splicing. Further studies to evaluate the isoforms of mRNAs in individuals with missense mutations in ESE sites would help to confirm the role of missense mutations in disease causation and to provide clinical insight into the significance of these mutations.

Acknowledgments

This research was supported, in part, by a cancer-prevention fellowship supported by National Cancer Institute grants R25 CA57730 (Robert M. Chamberlain, Ph.D., principal investi-

gator), CA70759 (to M.L.F.), CA16672, and National Human Genome Research Institute grant HG02275.

Electronic-Database Information

URLs for data presented herein are as follows:

ESEfinder, http://exon.cshl.org/ESE/index.html
Mutation Database on Pathogenic Mutations and Polymorphism for the *hMLH1* gene, http://www.nfdht.nl/database/database-mlh1.htm
Mutation Database on Pathogenic Mutations and Polymorphism for the *hMSH2* gene, http://www.nfdht.nl/database2/database-msh2.htm
Mutation Database Polymorphism for the *hMLH1* gene, http://www.nfdht.nl/MLH1-poly/poly-mlh1.htm
Mutation Database Polymorphism for the *hMLH2* gene, http://www.nfdht.nl/MSH2-poly/poly-msh2.htm
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for HNPCC)

References

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9:237–247

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25:106–110

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Cartegni L, Krainer AR (2002) Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nat Genet 30:377–384

Chabot B, LeBel C, Hutchison S, Nasim FH, Simard MJ (2003) Heterogeneous nuclear ribonucleoprotein particle A/B proteins and the control of alternative splicing of the mammalian heterogeneous nuclear ribonucleoprotein particle A1 pre-mRNA. Prog Mol Subcell Biol 31:59–88

Dagan T, Talmor Y, Graur D (2002) Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol Biol Evol 19:1022–1025

Fackenthal JD, Cartegni L, Krainer AR, Olopade OI (2002) *BRCA2* T2722R is a deleterious allele that causes exon skipping. Am J Hum Genet 71:625–631

Graveley BR (2000) Sorting out the complexity of SR protein functions. RNA 6:1197–1211

Hastings ML, Krainer AR (2001) Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol 13:302–309

Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol Cell Biol 20:1063–1071

Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12:1998–2012

Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM (2002) Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 19:149–164

Moseley CT, Mullis PE, Prince MA, Phillips JA 3rd (2002) An exon splice enhancer mutation causes autosomal dominant GH deficiency. J Clin Endocrinol Metab 87:847–852

Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 19:607–614

Peltomaki P, Vasen HF (1997) Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer. Gastroenterology 113:1146–1158

Pollard AJ, Krainer AR, Robson SC, Europe-Finner GN (2002) Alternative splicing of the adenylyl cyclase stimulatory G-protein G α(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3′-splice site. J Biol Chem 277:15241–15251

Stojdl DF, Bell JC (1999) SR protein kinases: the splice of life. Biochem Cell Biol 77:293–298