An MLH1 haplotype is over-represented on chromosomes carrying an HNPCC predisposing mutation in MLH1

P Hutter, J Wijnen, C Rey-Berthod, J Thiffault, P Verkuijlen, D Farber, N Hamel, B Bapat, S N Thibodeau, J Burn, J Wu, E MacNamara, K Heinimann, G Chong, W D Foulkes

Background: The mismatch repair gene, MLH1, appears to occur as two main haplotypes at least in white populations. These are referred to as A and G types with reference to the A/G polymorphism at IVS14-19. On the basis of preliminary experimental data, we hypothesised that deviations from the expected frequency of these two haplotypes could exist in carriers of disease associated MLH1 germ-line mutations.

Methods: We assembled a series (n=119) of germline MLH1 mutation carriers in whom phase between the haplotype and the mutation had been conclusively established. Controls, without cancer, were obtained from each contributing centre. Cases and controls were genotyped for the polymorphism in IVS14.

Results: Overall, 66 of 119 MLH1 mutations occurred on a G haplotype (55.5%), compared with 315 G haplotypes on 804 control chromosomes (39.2%, p=0.001). The odds ratio (OR) of a mutation occurring on a G rather than an A haplotype was 1.93 (95% CI 1.29 to 2.91). When we compared the haplotype frequencies in mutation bearing chromosomes carried by people of different nationalities with those seen in pooled controls, all groups showed a ratio of A/G haplotypes that was skewed towards G, except the Dutch group. On further analysis of the type of each mutation, it was notable that, compared with control frequencies, deletion and substitution mutations were preferentially represented on the G haplotype (p=0.003 and 0.005, respectively).

Conclusion: We have found that disease associated mutations in MLH1 appear to occur more often on one of only two known ancient haplotypes. The underlying reason for this observation is obscure, but it is tempting to suggest a possible role of either distant regulatory sequences or of chromatin structure influencing access to DNA sequence. Alternatively, differential behaviour of otherwise similar haplotypes should be considered as prime areas for further study.

H
ereditary non-polyposis colorectal cancer (HNPCC) is a dominant syndrome that affects about 1 in 1000 people. Patients with HNPCC have a family history of colorectal cancer at an early age, clinically characterised by a predominance of tumours in the proximal colon, a high frequency of synchronous and metachronous colorectal cancers, and an association with a variety of extra-colorectal tumours. Two to 5% of colon cancers are associated with a germline mutation in one of five mismatch repair (MMR) genes (MLH1, MSH2, MSH6, PMS2, MLH3). The MMR system has evolved to correct biosynthetic errors such as nucleotide misincorporations or misalignments during DNA replication. Together MLH1 and MSH2 account for at least 60% of all germline mutations found in families in which clinical diagnosis is based on the Amsterdam criteria. According to the database of the HNPCC consortium (http://www.nfhd.nl/database/mdbchoice.html), 17 mutations have been reported in MLH1 and 133 in MSH2. Defects in MLH1 also account for the majority of sporadic cancers exhibiting a characteristic tumour signature of DNA microsatellite instability, which are observed in a fraction of tumours of the colon, the endometrium, and the stomach. Between 8 and 12% of all colon cancers are associated with promoter hypermethylation of both MLH1 alleles at the somatic level. In addition to its role in DNA editing, the MLH1 gene has been shown to participate in mitotic and meiotic recombination, where it plays a role in the correction of heteroduplex and in apoptosis. In a previous study, we reported that two major haplotypes of the MLH1 gene appear to segregate in white populations, typically carrying either an A or a G nucleotide at IVS14-19. In particular, we observed that among 151 chromosomes stemming from nine European countries that carried a G at IVS14, all 151 had 11 CA dinucleotides at microsatellite D3S1611. In contrast with this, the 192 chromosomes which had an A at IVS14-19 exhibited six sizes of CA repeats, ranging from eight to 17, but never of 11 CAs. Marker D3S1611 is separated from IVS14-19 by 15 013 bp of genomic DNA, most of which corresponds to the large 11 253 bp long intron 13. Therefore, chromosomes carrying A or G at IVS14-19 will hereafter be referred to as A and G haplotypes, respectively. Preliminary results from 19 different MLH1 germline mutations had suggested that chromosomes carrying the G haplotype may more often harbour an MLH1 mutation causing HNPCC than chromosomes carrying the A haplotype, although the latter chromosomes are more abundant in the general population. In the present study, we have extended the analysis of the association between the A/G polymorphism at IVS14-19 to a series of 119 HNPCC associated MLH1 mutations stemming from Europe and North America.

MATERIALS AND METHODS

Polymorphism genotyping

In Switzerland, the A/G single nucleotide polymorphism at IVS14-19 of MLH1 was genotyped by direct sequence analysis.
of PCR products from genomic DNA. Genomic DNA was puri-
ified using standard phenol-chloroform extraction methods
and amplified for 30 cycles in 20 µl (3-5 pmol of each primer,
50 pmol/l of each dNTP, and 0.1 U of EXTRA-POL II DNA
polymerase from Eurobio, Les Ulis, France). Primers used
to genotype the A/G polymorphism were 5′-ATTTGCGCAACT-
GGTTGTA 3′ (forward primer) and 5′-TCAGTGAAATGATCA
GAAGTG 3′ (reverse primer). The reverse primer was tailed
with the M13 universal (-21) sequence, in order to use this
primer labelled with IRD800 dye (Lincoln, NE) for sequencing.
Cycle sequencing was performed using ThermoSequenase
(Nycomed Amersham, Buckinghamshire, UK). Amplicons
were usually sequenced in one direction only but occasionally
the results were confirmed by bidirectional sequencing. One µl
of reaction product was denatured and electrophoresed on
denaturing polyacrylamide gels (obtained by mixing 19 ml of
8% Sequagel XR from National Diagnostics, Atlanta, GA, with
952 µl of Long Ranger, FMC, Rockland, ME). Samples were
separated and analysed on a Li-Cor 4000 automated se-
duencer.

In Montreal, the A/G polymorphism was typed by direct
sequencing using the same primers indicated above and a
Cy5.5 labelled M13 primer, but using the Visible Genetics
apparatus (Toronto, ON) for sequence analysis. A PCR-RFLP
technique was also used as a rapid method of mutation detec-
tion for a subset of the controls. For this purpose the forward
primer 5′-TCTTTCAGCTCTGCTGCGCCT-3′ and the reverse primer
5′-ATAATAGAGAAGCTAAGTTAAAC-3′ were used. Following
amplification, the PCR products were digested with MaeII
(Roche Diagnostics, Mannheim, Germany). The products were
run on 8% polyacrylamide gels. The A allele was indicated by
a lower band (53 bp), whereas the G allele was represented by
a higher band (63 bp). In The Netherlands, genotyping was performed both by
denaturing gradient gel electrophoresis (DGGE) of exon 14
and by DNA sequencing as described previously.36

Statistical analysis
Cases and controls were compared using several statistical
methods. First, we compared the frequency of the two hap-
lootypes in the entire set of cases and controls using Fisher's
exact test (FET). Secondly, we calculated the odds ratio for the
association between carrying a germline MLH1 mutation and
the presence of the G haplotype, with 95% confidence
intervals (CI) estimated according to the method of Gart. To
evaluate if the effect observed was not entirely because of one
of the subsets of cases and controls, we used the Mantel-
Haenszel (M-H) method of establishing a pooled odds ratio
different sets of data. We also calculated Q, which
provides a p value for the probability that the subsets of data
analysed are derived from a single large population. A small p
value (<0.05) implies that the data should not be combined,
as they are not likely to be derived from a single population.
Finally, we recalculated the p values for each subgroup when
compared with all controls. For all statistical tests, the Arcus
Quickstat package was used (Addison Wesley Longman, UK).

Sources of the chromosomes studied
Chromosomes carrying MLH1 mutations and control chromo-
somes were ascertained from several geographical areas, as
shown in table 1. Ethnicity of the mutation carriers was not
precisely established, but our sampling is likely to reflect
the population from which the cases were drawn. Overall, MLH1
mutation carriers from North American centres will probably
be from more varied ethnic backgrounds than those from The
Netherlands or Switzerland. For this reason, we established
control populations from each centre, rather than using one
single source of control chromosomes. The country of origin of
mutation carriers and controls that were analysed at the three
coordinating centres were as follows: Leiden University Medi-
cal Centre, 62 cases, The Netherlands (n=23, 37%), USA
(n=13, 21%), UK (n=9, 15%), Germany (n=7, 11%), Ireland
(n=3, 5%), Norway (n=2, 3%) and one each from Portugal,
Denmark, Italy, Spain, and Australia, and 306 control chromo-
somes (all from The Netherlands); Sion and Basel Medical
Genetics Centres, 30 cases, Switzerland (n=21, 70%), Italy
(n=4, 13%), and one each from Germany, Portugal, Turkey,
Yugoslavia, and USA, and 216 control chromosomes (all from
Switzerland); McGill University. 27 cases, Montreal (n=11,
41%), Mayo Clinic, Minnesota (n=9, 33%), Mount Sinai Hos-
pital, Toronto (n=4, 15%), and UK (n=3, 11%), and 282 con-
trol chromosomes (all from Montreal). Following polymor-
phism analysis in one of the three centres, we reassigned cases
and controls on the basis of country/region of residence or
origin, where appropriate (tables 1 and 2). In each centre only
cases for which the phase could be established were included.
This resulted in some exclusions, as we were unable to deter-
mine phase in a further 12 MLH1 mutation carriers from
North America, five cases from Switzerland, and four cases
from The Netherlands. Among the 119 germline mutation
carriers, 98 were predicted or known to result in a truncated
protein (82.4%), whereas 21 (17.6%) were missense mutations
that have been shown to be disease causing, either on the basis
of segregation, or functional studies, or both. These 21 muta-
tions were as follows: C39R, S44F, Q62K, G67R (observed
three times), V77M, T117M, T117R, R265C (twice), N306K,
N551T, I565F, K618A (three times), P654L, L676P , and V716M
(twice). Mutations that were found more than once were all
found to have distinct haplotypes with respect to markers
closely linked to MLH1, as described by Hutter et al.36 No
controls had cancer. In The Netherlands, we included 240 popula-
tion control chromosomes and 66 chromosomes from spouses
of subjects carrying MLH1 mutations. As stated, for the North
American series all the controls came from Montreal. The eth-
nic distribution of controls was selected to reflect the cases in
this subgroup. The breakdown of the origin of the chromo-
somes from Montreal controls was as follows: 76 Ashkenazi
(27%), 68 British (24%), 56 southern European (20%), 50
other European (18%), and 30 French Canadian (11%) (table
2).

RESULTS
Overall, the G allele at IVS14-19 was over-represented on
chromosomes carrying a germline MLH1 mutation, compared
with controls (52.2% vs 39.2%, p=0.001, tables 1-3). The
association can also be expressed as an odds ratio (OR): the G
haplotype was 1.93 times more likely to be associated with a

<table>
<thead>
<tr>
<th>Country of residence</th>
<th>Total</th>
<th>A haplotype (%)</th>
<th>G haplotype (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Netherlands</td>
<td>23</td>
<td>14 (60.9)</td>
<td>9 (39.1)</td>
</tr>
<tr>
<td>USA</td>
<td>21</td>
<td>13 (47.8)</td>
<td>12 (52.2)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>21</td>
<td>7 (33.3)</td>
<td>14 (66.7)</td>
</tr>
<tr>
<td>Canada</td>
<td>15</td>
<td>6 (40.0)</td>
<td>9 (60.0)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>12</td>
<td>4 (33.3)</td>
<td>8 (66.7)</td>
</tr>
<tr>
<td>Germany</td>
<td>8</td>
<td>3 (37.5)</td>
<td>5 (62.5)</td>
</tr>
<tr>
<td>Others</td>
<td>17</td>
<td>8 (47.1)</td>
<td>9 (52.9)</td>
</tr>
<tr>
<td>Italy</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td>4</td>
<td>2 (50.0)</td>
<td>2 (50.0)</td>
</tr>
<tr>
<td>Norway</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portugal</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>119</td>
<td>53 (44.5)</td>
<td>66 (55.5)</td>
</tr>
</tbody>
</table>
Table 2 Haplotype frequencies in controls: ethnicity/country of origin of control chromosomes

<table>
<thead>
<tr>
<th>Country of residence</th>
<th>Total</th>
<th>A haplotype (%)</th>
<th>G haplotype (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Netherlands</td>
<td>306</td>
<td>187 (61.1)</td>
<td>119 (38.9)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>216</td>
<td>132 (61.1)</td>
<td>84 (38.9)</td>
</tr>
<tr>
<td>Canada*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashkenazi Jewish</td>
<td>76</td>
<td>46 (60.5)</td>
<td>30 (39.5)</td>
</tr>
<tr>
<td>British</td>
<td>68</td>
<td>47 (69.1)</td>
<td>21 (30.9)</td>
</tr>
<tr>
<td>French/French Canadian</td>
<td>30</td>
<td>17 (56.6)</td>
<td>13 (43.3)</td>
</tr>
<tr>
<td>Other European†</td>
<td>108</td>
<td>60 (55.6)</td>
<td>48 (44.4)</td>
</tr>
<tr>
<td>Canada (total)</td>
<td>282</td>
<td>170 (60.2)</td>
<td>112 (39.7)</td>
</tr>
<tr>
<td>Totals</td>
<td>804</td>
<td>489 (60.8)</td>
<td>315 (39.2)</td>
</tr>
</tbody>
</table>

*Canadian controls, ethnicity/country of origin was defined by birthplace of parents or grandparents. Thus we were able to match Canadian controls to US and UK cases, as both these groups were likely to have western European origins. In the case of Ashkenazi Jewish (AJ) controls, birthplace was not included. Unless indicated as AJ, Canadian controls did not report AJ origins.
†“Other European” included southern Europeans (n=56), eastern Europeans (n=24), and mixed non-French, non-British European ancestry (n=18).

Table 3 Comparison of haplotype frequencies in cases and controls

<table>
<thead>
<tr>
<th>Subjects</th>
<th>A haplotype</th>
<th>G haplotype</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>489</td>
<td>315</td>
<td>804</td>
</tr>
<tr>
<td>Cases</td>
<td>53</td>
<td>66</td>
<td>119</td>
</tr>
<tr>
<td>OR (95% CI)</td>
<td>1.93 [1.29 to 2.91]</td>
<td></td>
<td>0.001*</td>
</tr>
</tbody>
</table>

*p value reflects acceptance or rejection of the null hypothesis that the subgroups under study are all derived from one much larger population. This analysis showed that this should not be rejected (Q=3.19, 2 df, p=0.20). Of relevance to this finding is that the haplotype frequencies for A and G were stable in the control populations, varying over 804

case/controls from the three centres from 60.2% to 61.1% for the A haplotype, a difference of only 0.9% (table 2). Moreover, when considering Ashkenazi Jewish controls from Montreal, the frequency of the two haplotypes was 60.5% A and 39.5% G, which further suggests that this haplotype frequency is remarkably stable in different populations in the New and Old World. Some of the differences observed in the smaller subdivisions of the control populations are likely to be random variations based on small sample sizes (table 2).

We also repeated the analyses after excluding all subjects carrying disease associated missense mutations in MLH1. This resulted in an OR of 1.75 (95% CI 1.13 to 2.74, p=0.009). These findings do not suggest that the association we have identified is limited to any one type of mutation. We also studied the effect of the position of the mutation along the coding sequence of MLH1 on the frequency of the G haplotype. There was no evidence that the probability of the mutation occurring on the G haplotype varied according to the exon. This finding supports this hypothesis. When comparing the distribution of A/G haplotypes among chromosomes bearing deletions, insertions, and substitutions, the direction of the association between G haplotypes and mutation bearing chromosomes is positive for deletions and substitutions (ORs 2.9 and 2.1 respectively) and negative for insertions (OR 0.31), but because of the small sample sizes, these differences are not statistically significant. When compared with controls, deletion (n=37) and substitution mutations (n=70) were preferentially represented on the G haplotype (p=0.003 and 0.005 respectively). No significant difference was seen in the distribution of haplotypes for insertions (p=0.14), but only 12...
When considering only MMR, at least four speculative interpretations of our findings could be proposed. First, one or more of the above polymorphisms at BAT-21, D3S1611, and IVS14-19 may not be completely neutral with respect to MLH1 DNA repair function, and thus result in A/A genotypes being better than G/G genotypes at repairing replication errors occurring during the S phase of the cell cycle. This could be reminiscent of the common polymorphism N372H in the BRCA2 gene, which confers an increased risk of breast cancer, but clearly has other, unrelated functions. Secondly, more of the above three polymorphisms may be associated with MLH1 chromatin structure characteristics that render the MLH1 G haplotype slightly less accessible to DNA interacting proteins. This could have an effect, for instance, on the interaction with the MSH6 protein which is the actual mismatch recognition protein of the MSH2/MSH6 complex, thus altering MMR efficiency. If this were the case, both neutral polymorphisms and pathogenic mutations would be expected to occur more often on G than A haplotypes, and this effect should be restricted to MLH1. Thirdly, a slight functional difference in MLH1 between A and G alleles might be related to a MLH1 linked factor, located some distance apart from the three neutral polymorphisms, but capable of influencing MLH1 expression. In mice, distant enhancers of promoters have been identified some 250 kb away from the Hox gene complex. In Drosophila, long range transcription activators of genes have been found to be mediated by proteins which bind remote enhancer sequences, located several kb from the promoter. Between enhancers and promoters, insulator sequences have been identified, which interfere with enhancer-promoter communication. Allelic variants at such insulators can alter the physical proximity between enhancer and promoters, and a promoter on one chromosome can even be activated by an enhancer on the paired homologue. If similar factors were involved in A/G MLH1 variants, one would have to argue that an A/G promoter should be altered in order to avoid interference with the other functions of MLH1. For example, MLH1 protein can be produced alone in a baculovirus system, whereas PMS2 cannot be produced in the absence of MLH1 (G Marra, personal communication). Our previous results showed that the G alleles are consistently found on chromosomes that also differ from those carrying A alleles, at least with respect to nearby markers D3S1611 in IVS12 and BAT-21 in IVS11. On G alleles, D3S1611 monomorphic marker always carries 11 CA dinucleotides, whereas for the A allele, at least six sizes ranging from eight to 17 (but never 11) CA are found. BAT-21 marker comprises a run of 11 TA dinucleotides directly followed by a run of 21 T nucleotides, just 7 nt upstream of the acceptor splice junction of exon 12. This repeat is typically shortened by eight nucleotides on G alleles only.

**ACKNOWLEDGEMENTS**

We are very grateful to clinicians, patients, and their families who contributed to this study. P Hutter was supported by the following institutions: Recherche Suisse Contre le Cancer (RA346), Fondation

www.jmedgenet.com
Over-representation of MLH1 haplotype in HNPCC

Authors’ affiliations
P Hutter, C Rey-Berthod, Unit of Genetics, Institut Central des Hôpitaux Valaisans, Sion, Switzerland
J Wijnen, P Verkuiljen, Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands
N Hamel, W D Foulkes, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
D J Burn, S N Thibodeau, Sinai Hospital, Toronto, Canada
Davis-Jewish General Hospital, McGill University, Montreal, Quebec,
If you have a burning desire to respond to a paper published in Journal of Medical Genetics, why not make use of our “rapid response” option? Log on to our website (www.jmedgenet.com), find the paper that interests you, and send your response via email by clicking on the “eletters” option in the box at the top right hand corner. Providing it isn’t libellous or obscene, it will be posted within seven days. You can retrieve it by clicking on “read eletters” on our homepage. The editors will decide as before whether to publish it in a future paper issue as well as.
www.jmedgenet.com