

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Thompson, BA;Goldgar, DE;Paterson, C;Clendenning, M;Walters, R;Arnold, S;Parsons, MT;Michael, WD;Gallinger, S;Haile, RW;Hopper, JL;Jenkins, MA;Lemarchand, L;Lindor, NM;Newcomb, PA;Thibodeau, SN;Young, JP;Buchanan, DD;Tavtigian, SV;Spurdle, AB

Title:

A Multifactorial Likelihood Model for MMR Gene Variant Classification Incorporating Probabilities Based on Sequence Bioinformatics and Tumor Characteristics: A Report from the Colon Cancer Family Registry

Date:

2013-01-01

Citation:

Thompson, B. A., Goldgar, D. E., Paterson, C., Clendenning, M., Walters, R., Arnold, S., Parsons, M. T., Michael, W. D., Gallinger, S., Haile, R. W., Hopper, J. L., Jenkins, M. A., Lemarchand, L., Lindor, N. M., Newcomb, P. A., Thibodeau, S. N., Young, J. P., Buchanan, D. D., Tavtigian, S. V. & Spurdle, A. B. (2013). A Multifactorial Likelihood Model for MMR Gene Variant Classification Incorporating Probabilities Based on Sequence Bioinformatics and Tumor Characteristics: A Report from the Colon Cancer Family Registry. *Human Mutation*, 34 (1), pp.200-209. <https://doi.org/10.1002/humu.22213>.

Persistent Link:

<https://hdl.handle.net/11343/55539>

Supporting Information for this preprint is available from the
Human Mutation editorial office upon request (humu@wiley.com)

A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry

Bryony A. Thompson^{1,2*}, David E. Goldgar³, Carol Paterson¹, Mark Clendenning¹, Rhiannon Walters¹, Sven Arnold¹, Michael T. Parsons¹, Michael D. Walsh¹, Steven Gallinger⁴, Robert W. Haile⁵, John L. Hopper⁶, Mark A. Jenkins⁶, Loic LeMarchand⁷, Noralane M. Lindor⁸, Polly A. Newcomb⁹, Stephen N. Thibodeau¹⁰, Colon Cancer Family Registry, Joanne P. Young¹, Daniel D. Buchanan¹, Sean V. Tavtigian¹¹ and Amanda B. Spurdle¹

© 2012 Wiley Periodicals, Inc.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as [DOI: 10.1002/humu.22213].

- 1) Department of Genetics and Population Health, Queensland Institute of Medical Research, Herston, Brisbane, Australia.
- 2) School of Medicine, University of Queensland, Brisbane, Australia.
- 3) Department of Dermatology, University of Utah School of Medicine, Salt Lake City, Utah.
- 4) Cancer Care Ontario, Department of Surgery, Familial Gastrointestinal Cancer Registry, University of Toronto, Toronto, Ontario, Canada.
- 5) Department of Preventive Medicine, University of Southern California, Los Angeles, California.
- 6) Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Australia.
- 7) Cancer Research Center of Hawaii, University of Hawaii at Manoa, Honolulu, Hawaii.
- 8) Department of Health Science Research, Mayo Clinic Arizona, Scottsdale, Arizona.
- 9) Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington.
- 10) Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota.
- 11) Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah.

* Correspondence to: Bryony A. Thompson, Department of Genetics and Population Health, Queensland Institute of Medical Research, Locked Bag 2000, Royal Brisbane Hospital, Herston, QLD, 4029, Australia

Email: Bryony.Thompson@qimr.edu.au

Grant support: NHMRC ID 496616, Cancer Australia 1010859, NIH National Cancer Institute

grants RFA CA-95-011, U01 CA097735 (Australasian Colorectal Cancer Family Registry), U01 CA074799 (USC Familial Colorectal Neoplasia Collaborative Group), U01 CA074800 (Mayo Clinic Cooperative Family Registry for Colon Cancer Studies), U01 CA074783 (Ontario Registry for Studies of Familial Colorectal Cancer), U01 CA074794 (Seattle Colorectal Cancer Family Registry), U01 CA074806 (University of Hawaii Colorectal Cancer Family Registry), U01 CA078296 (University of California, Irvine Informatics Center).

Abstract

Mismatch repair (MMR) gene sequence variants of uncertain clinical significance are often identified in suspected Lynch syndrome families, and this constitutes a challenge for both researchers and clinicians. Multifactorial likelihood model approaches provide a quantitative measure of MMR variant pathogenicity, but first require input of likelihood ratios (LRs) for different MMR variation-associated characteristics from appropriate, well-characterized reference datasets. Microsatellite instability (MSI) and somatic *BRAF* tumor data for unselected colorectal cancer probands of known pathogenic variant status were used to derive LRs for tumor characteristics using the Colon Cancer Family Registry (CFR) resource. These tumor LRs were combined with variant segregation within families, and estimates of prior probability of pathogenicity based on sequence conservation and position, to analyze 44 unclassified variants identified initially in Australasian Colon CFR families. In addition, *in vitro* splicing analyses were conducted on the subset of variants based on bioinformatic splicing predictions. The LR in favor of pathogenicity was estimated to be ~12-fold for a colorectal tumor with a *BRAF* mutation-negative MSI-H phenotype. For 31 of the 44 variants, the posterior probabilities of pathogenicity were such that altered clinical management would be indicated. Our findings provide a working multifactorial likelihood model for classification that carefully considers mode of ascertainment for gene testing.

Key Words: unclassified variants; mismatch repair; multifactorial likelihood model; *BRAF*; MSI

Introduction

Lynch syndrome is an autosomal dominantly inherited disorder of predominantly colorectal and endometrial cancer predisposition caused by germline pathogenic variants in the DNA mismatch repair (MMR) genes *MLH1* (MIM# 120436), *MSH2* (MIM# 609309), *MSH6* (MIM# 600678), and to a lesser degree *PMS2* (MIM# 600259) (Lynch, et al., 2009; Viel, et al., 1998; Wang, et al., 1999). These pathogenic variants cause a functional defect in the DNA MMR complex, which leads to high levels of DNA microsatellite instability (MSI). Tumor MSI testing, and/or MMR protein immunostaining, is commonly used to identify potential Lynch syndrome cases (Aaltonen, et al., 1998; Lindor, et al., 2002; Salovaara, et al., 2000; Terdiman, et al., 2001). However, 10 to 15% of colorectal tumors detected in the general population have an MSI-H phenotype (Cunningham, et al., 1998; Herman, et al., 1998; Miyakura, et al., 2001; Toyota, et al., 1999), with molecular evidence indicating that these are largely caused by epigenetic silencing of the *MLH1* promoter (Deng, et al., 2004; Domingo, et al., 2004; Domingo, et al., 2005; Lubomierski, et al., 2005; McGivern, et al., 2004; Wang, et al., 2003; Young, et al., 2005). A single somatic mutation in the oncogene *BRAF* (MIM# 164757, *BRAF* c.1799T>A p.(Val600Glu)) is frequently present in MSI-H cancers with *MLH1* methylation (Davies, et al., 2002; Rajagopalan, et al., 2002), but not in MSI-H cancers arising in patients with a germline MMR gene pathogenic variant (Deng, et al., 2004; Domingo, et al., 2005; Lubomierski, et al., 2005; McGivern, et al., 2004; Wang, et al., 2003). Therefore *BRAF* p.Val600Glu has potential as a molecular marker to help distinguish MSI-H colorectal cancers caused by germline MMR gene pathogenic variant.

Between 20-50% of MMR gene sequence variants found by germline genetic testing in suspected Lynch syndrome colorectal and endometrial cancer cases are reported to be of uncertain clinical significance (Peltomaki and Vasen, 2004; Woods, et al., 2007). Establishing the pathogenicity of unclassified variants has direct clinical relevance, since pathogenic variant status determines clinical management with respect to genetic counselling approaches, presymptomatic screening, and choice/timing of possible prophylactic surgery for carriers.

Multifactorial likelihood analysis (also termed Bayesian Integrated evaluation) was established to address the issue of unclassified variants in the breast cancer susceptibility genes *BRCA1* (MIM# 113705) and *BRCA2* (MIM# 600185) (Goldgar, et al., 2008; Goldgar, et al., 2004). This approach provides a quantitative measure of variant pathogenicity, but first requires input of likelihood ratios (LRs) for different pathogenic variant-associated characteristics from appropriate, well-characterized reference datasets, with each LR comparing the probability of the observed data under the hypothesis that the variant is pathogenic compared to the corresponding probability that the variant is benign with respect to cancer risk. This method has since been proposed to assess cancer risk associated with rare variants of uncertain clinical significance for other cancer gene syndromes (Goldgar, et al., 2008; Tavtigian, et al., 2008), particularly since it provides a measure of causality in the form of the estimated posterior probability that a variant is pathogenic that can then be categorised into a qualitative five class scheme that is linked to clinical recommendations (Plon, et al., 2008).

We and others have proposed approaches to apply the multifactorial likelihood model to classification of MMR gene variants (Couch, et al., 2008; Hofstra, et al., 2008; Spurdle, 2010),

and have developed methodology to estimate the likelihood of pathogenicity based on variant segregation with disease in families (Arnold, et al., 2009). The extension of the multifactorial approach to include other features was recently reported in a study assessing 35 unclassified variants identified in Italian clinic-based families (Pastrello, et al., 2011). However, the stringent selection criteria for MMR gene sequencing of the reference sets used for derivation of LR estimates for tumor characteristics by Pastrello et al suggests that the LR estimates for tumor characteristics assigned in this Italian study were likely upwardly biased.

In this study, we have used a detailed set of qualitative criteria to consistently assign pathogenicity of MMR gene variants for a large set of colorectal cancer (CRC) cases (probands) with well-documented selection criteria for ascertainment and molecular characterization. We then compared the tumor MSI and *BRAF* mutation characteristics of probands with and without pathogenic MMR gene variants to estimate tumor LRs. Separate LRs were derived based on ascertainment criteria (i.e. clinic or population-based) to provide unbiased predictors of pathogenicity for inclusion in the multifactorial model, and were only applied to the relevant group. These tumor LRs were used together with existing approaches to assess variant segregation with disease in families, and estimates of prior probability of pathogenicity based on sequence conservation and position (see accompanying manuscript), to analyze 54 sequence variants in *MLH1*, *MSH2* and *MSH6* from 162 families. In addition, *in vitro* splicing analysis was performed for the subset of variants for which *in silico* analyses predicted a potential to cause splicing aberrations. We demonstrate the utility of both multifactorial likelihood and splicing analyses to classify exonic and intronic variants of MMR genes.

Methods and Materials

Subjects used in this study were recruited as colorectal cancer probands into the Colon Cancer Family Registry (Colon CFR) (Newcomb, et al., 2007). This includes probands providing data for reference sets (outlined below), and families analyzed for unclassified variant evaluation. Written informed consent was obtained from all participants. Recruitment of subjects and this laboratory-based study have been approved by the Human Research Ethics Committee of participating Institutions.

Selection criteria and testing for germline MMR gene sequence variation in the Colon CFR probands is described in Figure 1. All MMR gene sequence variants lodged in the Colon CFR database as of August 2009 were converted to HGVS nomenclature using the following cDNA reference sequences: NM_000249.3 for *MLH1*, NM_000251.1 for *MSH2* and NM_000179.2 for *MSH6*. Nucleotide numbering reflects cDNA numbering with +1 corresponding to the A of the ATG translation initiation codon in the reference sequence, with the initiation codon as codon 1. All variants were re-evaluated based on standardised qualitative classification criteria (Supp. Table S1). Probands with Class 5 pathogenic variants, or probands with either Class 1 not pathogenic/no clinical significance sequence alterations or no known pathogenic variant after extensive gene screening were included in reference datasets used to estimate the likelihood ratios. Supp. Table S2 lists all the Class 5 and Class 1 MMR gene sequence variants and the rationale or publications providing supportive information for their classification.

Supp. Table S3 lists all variants selected for evaluation using multifactorial and splicing approaches. Variants denoted as Class 2 (likely not pathogenic), Class 3 (uncertain), or Class 4 (likely pathogenic) from patients ascertained from the Australasian Colorectal Cancer Family Registry (ACCFR) were selected for evaluation in this study. Additionally, variants reported in Arnold et al. (2009) classified on segregation and splicing data only were also re-evaluated using the multifactorial approach. Clinical information was collated for all relevant ACCFR families, and when data became available for this project, also for Colon CFR families ascertained from other sites of the Colon CFR (Mayo Clinic, Ontario Familial Colorectal Cancer Registry and Fred Hutchinson Cancer Research Center) that were reported to carry at least one of the selected variants. Additional tumor and segregation data was extracted from publications to aid in the classification of some variants. The final sample set encompassed a total of 54 variants in 162 Colon CFR families, plus information from 26 families identified from the literature (188 families in total). All variants have been submitted to the InSiGHT locus-specific database (www.insight-group.org).

Splicing analyses were conducted on RNA extracted from lymphoblastoid cell lines (LCLs) established from available lymphocytes using EBV transformation based on the principles of Sugden and Mark (Sugden and Mark, 1977). cDNAs from LCLs established from nine healthy Red Cross donors were used as controls for *in vitro* splicing analyses.

Tumor Characteristics and Correlation with Pathogenic Variant Status

Characterization of colorectal tumors from Colon CFR probands was carried out previously as part of core Colon CFR activities (Newcomb, et al., 2007). Briefly, formalin fixed, paraffin

embedded tissue sections were stained for MMR proteins MLH1, MSH2, MSH6, and PMS2 (Lindor, et al., 2002). Tumors were analyzed for MSI status using 10 microsatellite markers (BAT25, BAT26, BAT40, BAT34, D5S346, D17S250, ACTC, D18S55, D10S197, and MYCL), comparing to normal tissue as reference where possible, and classified according to number of markers demonstrating instability: MSI-H for ≥ 3 ; MSI-L for 1–2; MSS for 0 unstable markers (Lindor, et al., 2002). MSI data was only extracted from peer-reviewed journal articles if at least five of the above markers were tested. Somatic *BRAF* p.Val600Glu mutation testing was done in all available CRCs as described previously (Buchanan, et al., 2010). Relevant information up to August 2009 was extracted from the Colon CFR database for all CRC-affected probands and used to generate cross-tabulations by pathogenic variant status, and by proband ascertainment, as shown in Table 1. Statistical significance of differences in distribution between pathogenic variant carriers and non-carriers was assessed using a Pearson's chi-squared test using STATA. A likelihood ratio (LR) for tumor characteristics was derived for MSI status and *BRAF* mutation status data using the formula:

LR of causality for tumor category = (% pathogenic variant carriers)/(% non-carriers)

The likelihood ratios for all the tumor categories are shown in Table 1.

Note: Since MSI status was a criterion for MMR gene variant testing of population-based probands in the Colon CFR database, but *NOT* for clinic-based probands (see Figure 1) (Newcomb, et al., 2007), the MSI likelihood ratios for population-based probands were based on estimations of the number of pathogenic variant carriers in all population-based cases with tumor MSI data, as detailed in Supp. Table S4.

BRAF mutation status provided no additional predictive capacity over MSI status for MSS and MSI-L probands, and categories were collapsed accordingly, to calculate a *BRAF* likelihood ratio for the subset of probands with an MSI-H tumor. The distribution of *BRAF* mutations by MMR-mutation status differed between the clinic- and population-based sample sets, and separate *BRAF* estimates were thus generated and used for clinic-based versus population-based probands.

Estimation of Prior Probability of Pathogenicity

Missense substitutions: The derivation of prior probabilities for missense substitutions are described in detail elsewhere (accompanying manuscript). Briefly, sequence conservation at relevant positions in protein multiple sequence alignments of MLH1, MSH2 and MSH6 were assessed using a combination of MAPP (Stone and Sidow, 2005) and a custom version of PolyPhen-2.1 (Adzhubei, et al., 2010) (<http://hci-lovd.hci.utah.edu/>). The MAPP impact score and custom PolyPhen-2.1 output “pph2 probability” were used as continuous variables. A minimum and maximum prior probability of pathogenicity of 0.10 and 0.90, respectively, was set to ensure that additional sources of information were required to reach posterior probabilities that alter clinical management of patients with variants (i.e. Class 1, Class 2, Class 4, and Class 5) i.e. direct classification based on the prior probability alone was not possible. All synonymous substitutions were assigned a prior of 0.10. Prior probabilities were derived from calibration of the above programs based on analysis of a consistently classified and carefully annotated reference set of MMR missense alterations that excluded all missense alterations known to alter splicing.

Intronic substitutions: There is a relative paucity of published information on splicing assays for intronic MMR gene variants (Supp. Table S5). All MMR gene variants located in the donor or acceptor dinucleotides assayed *in vitro* have been reported to be associated with splicing aberrations. However, the majority of variants reported in the literature appear to have been selected for *in vitro* analysis on the basis of bioinformatic prediction and the potential for negative publication bias for variants with no associated splicing aberration cannot be discounted. In the absence of a large study that has assessed the pathogenicity of intronic MMR gene variants without selection bias, we thus conservatively assumed prior probabilities of pathogenicity previously calculated for *BRCA1* and *BRCA2* variants based on analysis of a large dataset (Easton, et al., 2007). Namely, variants located in canonical GT-AG dinucleotides have a prior probability of 0.96 and other intronic variants and alterations in the 3'UTR were assigned a prior probability of 0.26.

Multifactorial Likelihood Classification

Likelihood ratios for segregation were derived by Bayes factor analysis adapted from the method of Thompson et al (2003), as described previously (Arnold, et al., 2009). Penetrance estimates for *MLH1* and *MSH2* variants were derived from Quehenberger et al. (2005) and those for *MSH6* variants were derived from Baglietto et al. (2010). The variants *MLH1* c.790+10A>G, *MLH1* c.2146G>A (p.Val716Met), *MSH2* c.339G>A, *MSH2* c.*221G>T, *MSH2* c.*226A>G and *MSH6* c.1186C>G (p.Leu396Val) were present in multiple families and co-occurred with pathogenic variants in some probands. The bi-allelic families concerned were excluded from multifactorial analysis to ensure the segregation likelihood ratio for estimating causality of the variant would not be biased by the presence of a pathogenic variant in the family. Bayesian scores were derived

twice for those families in which two different variants of unknown clinical significance were observed *in trans*, analysing risk associated with each variant separately, under the assumption that only one variant was causal.

Likelihood ratios (LR) for tumor characteristics shown in Table 1 were applied as follows: clinic-based probands and all relatives with known MSI status were assigned a LR according to MSI phenotype and recruitment location. All probands and relatives with MSI-H status and known *BRAF* mutation status were assigned an LR according to *BRAF* mutation status and site of recruitment. The MSI and *BRAF* likelihood ratios were used for CRCs only.

For each variant under study, the individual likelihood ratios available from segregation and tumor characteristics were multiplied to calculate an overall multifactorial likelihood ratio, under the assumption that each factor was statistically independent. Bayes rule was then used to calculate a posterior probability that the variant was pathogenic from the multifactorial likelihood ratios and the prior probability. Variants were classified according to the 5 class IARC quantitative scheme (Plon, et al., 2008), based on the posterior probability.

Splicing assays and bioinformatic prediction of splicing aberrations

Variants were selected for splicing analysis based on lymphoblastoid cell line (LCL) or lymphocyte availability. LCL cycloheximide splicing assays, PCR sample preparation and sample sequencing were done using methods described previously (Whiley, et al., 2011). cDNA (synthesized using random hexamers – Sigma, St Louis MO USA) was used as template in PCR reactions with specific primers targeting the potential splice-sites (Supp. Table S6). PCR for

variants was performed under the conditions presented in Supp. Table S6. Each RT-PCR analysis included a set of 9 healthy Red Cross control LCLs (see methods). The long-range DNA polymerase Elongase (Invitrogen, Carlsbad CA USA) was used to test for intron inclusion in variant carriers. PCR products from *MLH1* c.588+1G>T, *MLH1* c.678-1G>C, *MLH1* c.791-1G>C, *MLH1* c.884G>A p.(Ser295Asn), *MLH1* c.1559-2A>T, *MSH2* c.645+1G>A and *MSH2* c.1387-9T>A were cloned using pGEM®-T Vector (Promega Madison, WI, USA) and verified by sequencing.

Results from splicing assays were compared to bioinformatic predictions of splicing aberrations. We utilized Human Splicing Finder version 2.4 (www.umd.be/HSF/), which evaluates splicing signals present in any human gene by using matrices to predict 5' and 3' splice sites and splice regulatory sites using different algorithms, including Human Splicing Finder matrices and MaxEntScan (Desmet, et al., 2009; Yeo and Burge, 2004). The coding DNA HGVS nomenclature was input into HSF for each variant. We determined the difference between variant and wild-type output scores as a proportion of wild-type scores for HSF matrices and MaxEntScan. NNSplice (http://www.fruitfly.org/seq_tools/splice.html) (Reese, et al., 1997) was also used to assess the effect of variants on mRNA splicing, with the exact sequence input consecutively for wild-type and variant sequences, and a default minimum score of 0.4 was set for both 5' and 3' splice sites. The qualitative classification criteria in Spurdle et al. (2008) for splicing assay interpretation was applied in some instances to classify variants if it improved on the multifactorial likelihood classification.

Results

In this study we assessed 54 variants using multifactorial analysis, with the addition of prior probabilities for missense variants and a likelihood ratio for tumor characteristics. We also carried out splicing assays on all probands with LCLs available for *in vitro* analysis and no previous reported splicing assay results, including alterations in the splice consensus dinucleotides (*MLH1* c.588+1G>T, *MLH1* c.791-1G>C, *MLH1* c.1559-2A>T and *MSH2* c.645+1G>A).

Multifactorial Likelihood Analysis

We were able to classify 31 variants as either not pathogenic (9 variants) or pathogenic (22 variants) in *MLH1*, *MSH2* and *MSH6* using the multifactorial likelihood model (Table 2; detailed information used to derive individual LRs is provided in Supp. Table S7). In most instances, the likelihood ratios derived from tumor characteristics provided more information towards the final classification than segregation likelihood ratios, with obvious discordance between tumor results and segregation for only one variant *MSH2* c.488T>G. There were an additional 6 variants considered likely not pathogenic and 4 likely pathogenic. Tumors from 3 cases that are carriers of an *MSH2* variant in the 3' untranslated region (*MSH2* c.*129T>C and *MSH2* c.*226A>G) demonstrated a tumor phenotype indicative of sporadic CRC (a positive mutation status for *BRAF* p.Val600Glu, MSI-H and *MLH1* protein loss), which is consistent with the neutral and likely neutral classifications observed.

The multifactorial classification, for each variant assessed in this study is reported on a newly created LOVD based MMR database (<http://hci-lovd.hci.utah.edu/>), which details the prior probabilities, likelihood ratios for the individual components, and posterior probabilities.

As shown in Table 2, there were 9 variants that were observed with known pathogenic variants, and 8 variants that were observed with another variant of unknown clinical significance. The earliest age of onset of a Lynch cancer for these probands was 37 years old (range 37 to 64), and none had an abnormal clinical phenotype, suggesting that the carriers of these variants were not carriers of two different *bona fide* pathogenic variants. Indeed, the final classifications from multifactorial analysis (which at this point in time does not incorporate a LR for co-occurrence) were Class 1 (4/9 variants), Class 2 (4/9 variants) and uncertain (1/9 variants) for variants co-observed with a known pathogenic variant.

***In vitro* splicing assays**

Several of the cases assayed expressed varying levels of naturally occurring alternative splicing isoforms, particularly *MLH1* Δ exon 9/10, *MLH1* Δ exon 3, a 227 base-pair insertion of intron 1 and 145 base-pair insertion of intron 2 both in *MLH1* (Figure 2a). Splicing aberrations not identified in controls were observed for 7 out of the 26 variants assayed (Supp. Table S8, Figure 2). The splicing programs predicted the interruption of the intron-exon junction for 5 variants, all of which showed aberrant splicing *in vitro*. The variants *MLH1* c.588+1G>T, *MLH1* c.791-1G>C, *MLH1* c.884G>A p.(Ser295Asn), and *MLH1* c.1559-2A>T all result in exon skipping (Figure 2a, b, c). *MSH2* c.645+1G>A produces multiple transcripts: an in-frame deletion of exon 3 and an out-of-frame product from cryptic donor site usage 154 base-pairs upstream of the wild-

type donor site (Figure 2d). These splicing results were consistent with multifactorial likelihood classifications for these variants, Class 4 (*MLH1* c.884G>A) or Class 5 (the remaining intron-exon junction variants). Moreover, analysis of cDNA sequence for the exonic variant *MLH1* c.884G>A indicated that the variant allele did not express any full length transcript, justifying that the classification for this variant may be upgraded to Class 5 on the basis of the splicing data.

Bioinformatic prediction of aberrant splicing was consistent for the remaining two variants displaying aberrant splicing. *MLH1* c.678-1G>C interrupts the wild-type acceptor splice site and creates a *de novo* acceptor leading to a 2 base-pair out-of-frame deletion identified by sequencing (Figure 2e), both of which were predicted by all 3 of the bioinformatic splicing programs (Supp. Table S8). *MSH2* c.1387-9T>A creates a *de novo* acceptor site causing a 7 base-pair inclusion of intron 8 (Figure 2f), with *de novo* acceptor creation and interruption of the wild-type acceptor site predicted by MaxEntScan (Supp. Table S8). Although further quantitative splicing analysis would be required to determine if the splicing aberrations caused by *MLH1* c.678-1G>C and *MSH2* c.1387-9T>A are associated with a minor or major transcript, the Class 4 likely pathogenic classification from multifactorial analysis for these variants suggests that these splicing aberrations are causative.

In addition, splicing assay results showed no evidence of an aberration for the intronic variant *MSH2* c.2005+8dupA, justifying that this variant can be altered from Class 3 uncertain (indicated by multifactorial analysis) to Class 2 likely not pathogenic. Splicing data did not alter the classifications of any other variants.

Discussion

In this study we derived likelihood ratios from the Colon CFR dataset for the CRC characteristics of MSI and *BRAF* mutation status for use in multifactorial analysis, carefully considering issues relating to ascertainment bias in this large and well-characterised dataset. We provide LRs that are applicable to variants discovered within clinic-based or population-based datasets, with the caveat that the reference pathogenic variants in this study are also representative of studies outside the Colon CFR dataset.

These LRs for tumor features were then used in a multifactorial model together with a segregation likelihood ratio and an estimated prior probability to assess clinical significance of 54 MMR gene variants, and contributed noticeably to classification. In total, application of the model altered the class of 31 variants under study (21 as Class 1 or 5, and another 10 Class 2 or 4), with implications for the future management of all families with these variants. Additionally, we were able to confirm the Class 5 classification for 10 variants considered pathogenic using largely qualitative information from Arnold et al. (2009). The priors based on *in silico* analysis of the 25 missense variants were in agreement with tumor and segregation data except for the variants MLH1 p.(Lys618Ala) and MSH2 p.(Ala636Pro). Underestimation of the prior for the latter Ashkenazi Jewish founder pathogenic variant traces to the observation that the position corresponding to Ala636 is quite variable in our protein multiple sequence alignments; moreover, mutS crystal structures do not place the residue in an alpha helix (Lamers, et al., 2000; Obmolova, et al., 2000). Consequently, the substitution to proline is not obviously destabilizing and neither MAPP nor PolyPhen gives the variant a high prior probability. Nonetheless, Foulkes

et al. (2002) suggested that the substitution could cause steric hindrance that might interrupt ATP hydrolysis. Still, in spite of the relatively low prior probability for this variant, the data from tumor characteristics and co-segregation were sufficient to provide a posterior probability indicative of pathogenicity. The classification of *MLH1* c.1852_1853delAAinsGC p.(Lys618Ala) is a contentious issue with some functional assays suggesting reduced protein function (Belvederesi, et al., 2006; Blasi, et al., 2006; Guerrette, et al., 1999; Kondo, et al., 2003; Perera and Bapat, 2008), but family studies (including this study) indicating it is not pathogenic (Castillejo, et al., 2011). Consistent with these inconclusive findings, a recent case-control study concluded that the variant was not fully penetrant, and was associated with a two-fold increase in risk of Lynch syndrome-associated tumors (Medeiros, et al., 2012).

There is currently no sufficiently large and well-characterised dataset to estimate prior probabilities of pathogenicity based on *in silico* splicing data for the MMR repair genes, and it is not feasible to estimate a splicing prior from the literature due to bias in selection of variants for *in vitro* mRNA assays (including this study with preselection by bioinformatic prediction splicing assays). Therefore, since the splicing machinery and regulatory elements are similar between genes (Baralle and Baralle, 2005), the *BRCA1/2* splicing prior probabilities (Easton, et al., 2007) were used in the multifactorial analysis as a conservative estimate of prior probability for all intronic variants (within and outside acceptor/donor dinucleotides). Our analysis of possible splicing aberrations for a subset of variants supported the findings from multifactorial analysis, with 9/12 demonstrating splicing aberrations having a posterior probability Class 5. The remaining 3 variants were Class 4 from multifactorial analysis, and semi-quantitative allelic expression analysis for one exonic variant indicated that it could be considered Class 5 using qualitative interpretation {Spurdle, 2008 #587}.

It was not appropriate to utilise other LRs derived using BRCA datasets such as co-occurrence and family history (Easton, et al., 2007; Goldgar, et al., 2004) due to the different phenotypes present for the mismatch repair genes compared to *BRCA1* and *BRCA2*. Although it is known that co-occurrence or co-observation of pathogenic variants in the MMR genes (also termed compound heterozygosity or bi-allelic mutations) is associated with very early onset gastrointestinal tumors (under 20 years old), haematologic malignancies and neurofibromatosis type 1-like features (Bandipalliam, 2005), further study is required to assign with confidence the specific features that can be used to derive LRs for pathogenic variants that co-occur (same MMR gene) or are co-observed (different MMR genes). Similarly the family history profiles associated with MMR gene pathogenic variant status will have to be assessed from large well-characterised datasets to derive LRs for this component.

This is the first instance of incorporating *BRAF* mutation status in the multifactorial model. Mining of the Colon CFR MSI and somatic *BRAF* tumor data from unselected colon cancer patients of known pathogenic variant status indicates that a *BRAF* mutation negative MSI-H colon tumor is associated with a ~12-fold increase in likelihood of pathogenicity. This data also suggests that MSI in a tumor is a better predictor of pathogenic variant carrier status than absence of a *BRAF* mutation. Positive *BRAF* mutation status was most informative in identifying probable non-pathogenic variants, as was MSS tumors. These findings are reflected by observations in the literature (Lucci-Cordisco, et al., 2006; Parsons, et al., 2012; Zaanani, et al., 2011). However, the incidence of MSI in sporadic CRCs (Cunningham, et al., 1998; Herman, et al., 1998; Miyakura, et al., 2001; Toyota, et al., 1999) highlights the importance of a

combinatorial tumor characteristic likelihood ratio. Using MSI alone can sometimes lead to over-estimation of pathogenicity. For the variant *MLH1* c.2146G>A p.(Val716Met) there are two MSI-H tumors eligible for analysis that have not had *BRAF* mutation testing, which increase the odds of causality. However, additional evidence suggests that this variant is not pathogenic and that something else other than the missense substitution is causing the MSI. A likelihood ratio based on *MLH1* promoter methylation would also be a beneficial addition to the multifactorial model as a negative indicator of Lynch syndrome.

BRAF mutation status is commonly associated with *MLH1* promoter hypermethylation in MSI-H sporadic CRCs (Deng, et al., 2004; McGivern, et al., 2004). Therefore, the presence of *BRAF* p.Val600Glu in a CRC should be sufficient to exclude MMR gene variant screening for Lynch syndrome (Bessa, et al., 2008; Bouzourene, et al., 2010; Jensen, et al., 2008; Loughrey, et al., 2007). However, there are seven pathogenic MMR variants reported in probands from this Colon CFR dataset and additional pathogenic variants reported in the literature (Lagerstedt Robinson, et al., 2007; Walsh, et al., 2009) that have *BRAF* mutation positive tumors, suggesting that recognition of *bona fide* Lynch tumors may be missed using such triaging for gene variation screening.

Few studies have addressed classifying MMR gene variants using multifactorial analysis. To our knowledge, three papers have tried integrating tumor characteristics with multiple lines of evidence to classify variants. Arnold et al. (2009) used Bayesian segregation analysis, but used tumor pathology data qualitatively together with other observational data (functional data and family history) and *in vitro* splicing data to classify variants. Barnetson et al. (2008) devised an

arbitrary scoring system for variant classification, that included IHC and MSI tumor status as points of evidence, but there was no attempt to derive an empirical LR. Furthermore, Pastrello et al. (2011) described a Bayesian approach to MMR variant classification that included a likelihood ratio for clinicopathologic features such as IHC and MSI based on a large tumor dataset (Engel, et al., 2006). Nevertheless, ascertainment bias in the dataset was not taken into account in derivation of the LRs, therefore leading to an overestimation in the likelihood of pathogenicity for variant carriers. In the estimation of the tumor characteristic LR we have taken into account ascertainment bias for the tumors, and chose not to estimate a likelihood ratio based on IHC data due to the high correlation between tumor protein loss and MSI, and because some missense substitutions can lead to stably expressed non-functional proteins not identified using IHC (Mangold, et al., 2005; van Riel, et al., 2010).

The tumor characteristics likelihood ratio based on MSI and *BRAF* mutation status was an integral component in the classification of nearly two thirds of the variants reported in this study. Therefore, this LR used in combination with segregation analysis and an *in silico* missense prior probability is a beneficial addition to the MMR multifactorial model for variant classification. While the strength of the multifactorial approach is the ability to draw on multiple data sources readily available in the clinical setting, we do acknowledge a number of assumptions that underlie the basic model. Namely, that the individual components are mutually independent, that all variants of clinical significance have similar cancer penetrances to known pathogenic variants, that estimates of penetrance used in segregation analysis have been accurately estimated, and that the phenotypic features of different types of pathogenic variants (missense,

splicing, truncating) are sufficiently similar to allow likelihood ratios estimates from one type of pathogenic variant to be relevant to other classes.

The results of the multifactorial analyses will affect the clinical management of 76% (41/54) of the variants reported within. Additionally, a mismatch repair gene public LOVD-based database has been created to report these classifications and future multifactorial likelihood classifications.

Acknowledgments

We thank the many families in the Colon Cancer Family Registry who have participated in the research programs, and the Australian Red Cross Blood Services (ARCBS) donors who participated as healthy controls in this study. We are grateful to Rachel Morris and the staff at ARCBS for their assistance with the collection of risk factor information and blood samples, and Melanie Higgins, Kimberley Hinze, Felicity Lose, and members of the Molecular Cancer Epidemiology Laboratory for their assistance with collection and processing of blood samples. We also acknowledge Russell Bell for his assistance in the set up of the LOVD website. B.A.T was awarded a PhD scholarship from the Cancer Council Queensland. A.B.S. is an NHMRC Senior Research Fellow. This work was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011 and through cooperative agreements with members of the Colon Cancer Family Registry and P.I.s. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR. None of the authors have a conflict of interest.

References

- Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomaki P, Chadwick RB, Kaariainen H, Eskelinen M, Jarvinen H and others. 1998. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med* 338:1481-7.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7:248-9.
- Arnold S, Buchanan DD, Barker M, Jaskowski L, Walsh MD, Birney G, Woods MO, Hopper JL, Jenkins MA, Brown MA and others. 2009. Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum Mutat* 30:757-70.
- Baglietto L, Lindor NM, Dowty JG, White DM, Wagner A, Gomez Garcia EB, Vriends AH, Cartwright NR, Barnetson RA, Farrington SM and others. 2010. Risks of Lynch syndrome cancers for MSH6 mutation carriers. *J Natl Cancer Inst* 102:193-201.
- Bandipalliam P. 2005. Syndrome of early onset colon cancers, hematologic malignancies & features of neurofibromatosis in HNPCC families with homozygous mismatch repair gene mutations. *Fam Cancer* 4:323-33.
- Baralle D, Baralle M. 2005. Splicing in action: assessing disease causing sequence changes. *J Med Genet* 42:737-48.
- Barnetson RA, Cartwright N, van Vliet A, Haq N, Drew K, Farrington S, Williams N, Warner J, Campbell H, Porteous ME and others. 2008. Classification of ambiguous mutations in

- DNA mismatch repair genes identified in a population-based study of colorectal cancer. *Hum Mutat* 29:367-74.
- Belvederesi L, Bianchi F, Loretelli C, Gagliardini D, Galizia E, Bracci R, Rosati S, Bearzi I, Viel A, Cellerino R and others. 2006. Assessing the pathogenicity of MLH1 missense mutations in patients with suspected hereditary nonpolyposis colorectal cancer: correlation with clinical, genetic and functional features. *Eur J Hum Genet* 14:853-9.
- Bessa X, Balleste B, Andreu M, Castells A, Bellosillo B, Balaguer F, Castellvi-Bel S, Paya A, Jover R, Alenda C and others. 2008. A prospective, multicenter, population-based study of BRAF mutational analysis for Lynch syndrome screening. *Clin Gastroenterol Hepatol* 6:206-14.
- Blasi MF, Ventura I, Aquilina G, Degan P, Bertario L, Bassi C, Radice P, Bignami M. 2006. A human cell-based assay to evaluate the effects of alterations in the MLH1 mismatch repair gene. *Cancer Res* 66:9036-44.
- Bouzourene H, Hutter P, Losi L, Martin P, Benhattar J. 2010. Selection of patients with germline MLH1 mutated Lynch syndrome by determination of MLH1 methylation and BRAF mutation. *Fam Cancer* 9:167-72.
- Buchanan DD, Sweet K, Drini M, Jenkins MA, Win AK, English DR, Walsh MD, Clendenning M, McKeone DM, Walters RJ and others. 2010. Risk factors for colorectal cancer in patients with multiple serrated polyps: a cross-sectional case series from genetics clinics. *PLoS One* 5:e11636.
- Caldes T, Godino J, de la Hoya M, Garcia Carbonero I, Perez Segura P, Eng C, Benito M, Diaz-Rubio E. 2002. Prevalence of germline mutations of MLH1 and MSH2 in hereditary nonpolyposis colorectal cancer families from Spain. *Int J Cancer* 98:774-9.

- Castillejo A, Guarinos C, Martinez-Canto A, Barbera VM, Egoavil C, Castillejo MI, Perez-Carbonell L, Sanchez-Heras AB, Segura A, Ochoa E and others. 2011. Evidence for classification of c.1852_1853AA>GC in MLH1 as a neutral variant for Lynch syndrome. *BMC Med Genet* 12:12.
- Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N. 2008. Assessment of functional effects of unclassified genetic variants. *Hum Mutat* 29:1314-26.
- Cunningham JM, Christensen ER, Tester DJ, Kim CY, Roche PC, Burgart LJ, Thibodeau SN. 1998. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 58:3455-60.
- Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W and others. 2002. Mutations of the BRAF gene in human cancer. *Nature* 417:949-54.
- Deng G, Bell I, Crawley S, Gum J, Terdiman JP, Allen BA, Truta B, Sleisenger MH, Kim YS. 2004. BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. *Clin Cancer Res* 10:191-5.
- Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37:e67.
- Domingo E, Laiho P, Ollikainen M, Pinto M, Wang L, French AJ, Westra J, Frebourg T, Espin E, Armengol M and others. 2004. BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing. *J Med Genet* 41:664-8.
- Domingo E, Niessen RC, Oliveira C, Alhopuro P, Moutinho C, Espin E, Armengol M, Sijmons RH, Kleibeuker JH, Seruca R and others. 2005. BRAF-V600E is not involved in the

- colorectal tumorigenesis of HNPCC in patients with functional MLH1 and MSH2 genes. *Oncogene* 24:3995-8.
- Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ and others. 2007. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet* 81:873-83.
- Engel C, Forberg J, Holinski-Feder E, Pagenstecher C, Plaschke J, Kloor M, Poremba C, Pox CP, Ruschoff J, Keller G and others. 2006. Novel strategy for optimal sequential application of clinical criteria, immunohistochemistry and microsatellite analysis in the diagnosis of hereditary nonpolyposis colorectal cancer. *Int J Cancer* 118:115-22.
- Foulkes WD, Thiffault I, Gruber SB, Horwitz M, Hamel N, Lee C, Shia J, Markowitz A, Figer A, Friedman E and others. 2002. The founder mutation MSH2*1906G-->C is an important cause of hereditary nonpolyposis colorectal cancer in the Ashkenazi Jewish population. *Am J Hum Genet* 71:1395-412.
- Gallinger S, Aronson M, Shayan K, Ratcliffe EM, Gerstle JT, Parkin PC, Rothenmund H, Croitoru M, Baumann E, Durie PR and others. 2004. Gastrointestinal cancers and neurofibromatosis type 1 features in children with a germline homozygous MLH1 mutation. *Gastroenterology* 126:576-85.
- Genuardi M, Anti M, Capozzi E, Leonardi F, Fornasarig M, Novella E, Bellacosa A, Valenti A, Gasbarrini GB, Roncucci L and others. 1998. MLH1 and MSH2 constitutional mutations in colorectal cancer families not meeting the standard criteria for hereditary nonpolyposis colorectal cancer. *Int J Cancer* 75:835-9.

- Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS. 2008. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum Mutat* 29:1265-72.
- Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am J Hum Genet* 75:535-44.
- Guerrette S, Acharya S, Fishel R. 1999. The interaction of the human MutL homologues in hereditary nonpolyposis colon cancer. *J Biol Chem* 274:6336-41.
- Herman JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa JP, Markowitz S, Willson JK, Hamilton SR, Kinzler KW and others. 1998. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A* 95:6870-5.
- Hofstra RM, Spurdle AB, Eccles D, Foulkes WD, de Wind N, Hoogerbrugge N, Hogervorst FB. 2008. Tumor characteristics as an analytic tool for classifying genetic variants of uncertain clinical significance. *Hum Mutat* 29:1292-303.
- Jensen LH, Lindebjerg J, Byriel L, Kolvraa S, Cruger DG. 2008. Strategy in clinical practice for classification of unselected colorectal tumours based on mismatch repair deficiency. *Colorectal Dis* 10:490-7.
- Kamory E, Tanyi M, Kolacsek O, Olasz L, Toth L, Damjanovich L, Csuka O. 2006. Two germline alterations in mismatch repair genes found in a HNPCC patient with poor family history. *Pathol Oncol Res* 12:228-33.
- Kondo E, Suzuki H, Horii A, Fukushige S. 2003. A yeast two-hybrid assay provides a simple way to evaluate the vast majority of hMLH1 germ-line mutations. *Cancer Res* 63:3302-8.

- Lagerstedt Robinson K, Liu T, Vandrovcova J, Halvarsson B, Clendenning M, Frebourg T, Papadopoulos N, Kinzler KW, Vogelstein B, Peltomaki P and others. 2007. Lynch syndrome (hereditary nonpolyposis colorectal cancer) diagnostics. *J Natl Cancer Inst* 99:291-9.
- Lamers MH, Perrakis A, Enzlin JH, Winterwerp HH, de Wind N, Sixma TK. 2000. The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. *Nature* 407:711-7.
- Lindor NM, Burgart LJ, Leontovich O, Goldberg RM, Cunningham JM, Sargent DJ, Walsh-Vockley C, Petersen GM, Walsh MD, Leggett BA and others. 2002. Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors. *J Clin Oncol* 20:1043-8.
- Loughrey MB, Waring PM, Tan A, Trivett M, Kovalenko S, Beshay V, Young MA, McArthur G, Boussioutas A, Dobrovic A. 2007. Incorporation of somatic BRAF mutation testing into an algorithm for the investigation of hereditary non-polyposis colorectal cancer. *Fam Cancer* 6:301-10.
- Lubomierski N, Plotz G, Wormek M, Engels K, Kriener S, Trojan J, Jungling B, Zeuzem S, Raedle J. 2005. BRAF mutations in colorectal carcinoma suggest two entities of microsatellite-unstable tumors. *Cancer* 104:952-61.
- Lucci-Cordisco E, Boccuto L, Neri G, Genuardi M. 2006. The use of microsatellite instability, immunohistochemistry and other variables in determining the clinical significance of MLH1 and MSH2 unclassified variants in Lynch syndrome. *Cancer Biomark* 2:11-27.

- Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR. 2009. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* 76:1-18.
- Mangold E, Pagenstecher C, Friedl W, Fischer HP, Merkelbach-Bruse S, Ohlendorf M, Friedrichs N, Aretz S, Buettner R, Propping P and others. 2005. Tumours from MSH2 mutation carriers show loss of MSH2 expression but many tumours from MLH1 mutation carriers exhibit weak positive MLH1 staining. *J Pathol* 207:385-95.
- McGivern A, Wynter CV, Whitehall VL, Kambara T, Spring KJ, Walsh MD, Barker MA, Arnold S, Simms LA, Leggett BA and others. 2004. Promoter hypermethylation frequency and BRAF mutations distinguish hereditary non-polyposis colon cancer from sporadic MSI-H colon cancer. *Fam Cancer* 3:101-7.
- Medeiros F, Lindor NM, Couch FJ, Highsmith WE, Jr. 2012. The germline MLH1 K618A variant and susceptibility to lynch syndrome-associated tumors. *J Mol Diagn* 14:264-73.
- Miyakura Y, Sugano K, Konishi F, Ichikawa A, Maekawa M, Shitoh K, Igarashi S, Kotake K, Koyama Y, Nagai H. 2001. Extensive methylation of hMLH1 promoter region predominates in proximal colon cancer with microsatellite instability. *Gastroenterology* 121:1300-9.
- Muller-Koch Y, Kopp R, Lohse P, Baretton G, Stoetzer A, Aust D, Daum J, Kerker B, Gross M, Dietmeier W and others. 2001. Sixteen rare sequence variants of the hMLH1 and hMSH2 genes found in a cohort of 254 suspected HNPCC (hereditary non-polyposis colorectal cancer) patients: mutations or polymorphisms? *Eur J Med Res* 6:473-82.
- Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, Hall D, Hopper JL, Jass J, Le Marchand L and others. 2007. Colon Cancer Family Registry: an international

- resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 16:2331-43.
- Obmolova G, Ban C, Hsieh P, Yang W. 2000. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* 407:703-10.
- Parsons MT, Buchanan DD, Thompson B, Young JP, Spurdle AB. 2012. Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. *Journal of Medical Genetics* 49:151-157.
- Pastrello C, Pin E, Marroni F, Bedin C, Fornasarig M, Tibiletti MG, Olini C, Ponz de Leon M, Urso ED, Della Puppa L and others. 2011. Integrated analysis of unclassified variants in mismatch repair genes. *Genet Med* 13:115-24.
- Peltomaki P, Vasen H. 2004. Mutations associated with HNPCC predisposition -- Update of ICG-HNPCC/INSiGHT mutation database. *Dis Markers* 20:269-76.
- Perera S, Bapat B. 2008. The MLH1 variants p.Arg265Cys and p.Lys618Ala affect protein stability while p.Leu749Gln affects heterodimer formation. *Hum Mutat* 29:332.
- Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* 29:1282-91.
- Quehenberger F, Vasen HF, van Houwelingen HC. 2005. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet* 42:491-6.

- Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. 2002. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. *Nature* 418:934.
- Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. *Journal of computational biology : a journal of computational molecular cell biology* 4:311-23.
- Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, Harkonen N, Julkunen R, Kangas E, Ojala S and others. 2000. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol* 18:2193-200.
- Spurdle AB. 2010. Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. *Curr Opin Genet Dev* 20:315-23.
- Spurdle AB, Couch FJ, Hogervorst FB, Radice P, Sinilnikova OM. 2008. Prediction and assessment of splicing alterations: implications for clinical testing. *Hum Mutat* 29:1304-13.
- Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 15:978-86.
- Sugden B, Mark W. 1977. Clonal transformation of adult human leukocytes by Epstein-Barr virus. *J Virol* 23:503-8.
- Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat* 29:1327-36.
- Terdiman JP, Gum JR, Jr., Conrad PG, Miller GA, Weinberg V, Crawley SC, Levin TR, Reeves C, Schmitt A, Hepburn M and others. 2001. Efficient detection of hereditary nonpolyposis colorectal cancer gene carriers by screening for tumor microsatellite instability before germline genetic testing. *Gastroenterology* 120:21-30.

- Thompson D, Easton DF, Goldgar DE. 2003. A full-likelihood method for the evaluation of causality of sequence variants from family data. *Am J Hum Genet* 73:652-5.
- Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. 1999. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* 96:8681-6.
- van Riel E, Ausems MG, Hogervorst FB, Kluijdt I, van Gijn ME, van Echtelt J, Scheidel-Jacobse K, Hennekam EF, Stulp RP, Vos YJ and others. 2010. A novel pathogenic MLH1 missense mutation, c.112A > C, p.Asn38His, in six families with Lynch syndrome. *Hered Cancer Clin Pract* 8:7.
- Viel A, Novella E, Genuardi M, Capozzi E, Fornasarig M, Pedroni M, Santarosa M, De Leon MP, Della Puppa L, Anti M and others. 1998. Lack of PMS2 gene-truncating mutations in patients with hereditary colorectal cancer. *Int J Oncol* 13:565-9.
- Walsh MD, Buchanan DD, Walters R, Roberts A, Arnold S, McKeone D, Clendenning M, Ruzskiewicz AR, Jenkins MA, Hopper JL and others. 2009. Analysis of families with Lynch syndrome complicated by advanced serrated neoplasia: the importance of pathology review and pedigree analysis. *Fam Cancer* 8:313-23.
- Wang L, Cunningham JM, Winters JL, Guenther JC, French AJ, Boardman LA, Burgart LJ, McDonnell SK, Schaid DJ, Thibodeau SN. 2003. BRAF mutations in colon cancer are not likely attributable to defective DNA mismatch repair. *Cancer Res* 63:5209-12.
- Wang Q, Lasset C, Desseigne F, Saurin JC, Maugard C, Navarro C, Ruano E, Descos L, Trillet-Lenoir V, Bosset JF and others. 1999. Prevalence of germline mutations of hMLH1, hMSH2, hPMS1, hPMS2, and hMSH6 genes in 75 French kindreds with nonpolyposis colorectal cancer. *Hum Genet* 105:79-85.

- Whiley PJ, Guidugli L, Walker LC, Healey S, Thompson BA, Lakhani SR, Da Silva LM, Investigators K, Tavtigian SV, Goldgar DE and others. 2011. Splicing and multifactorial analysis of intronic BRCA1 and BRCA2 sequence variants identifies clinically significant splicing aberrations up to 12 nucleotides from the intron/exon boundary. *Hum Mutat* 32:678-87.
- Woods MO, Williams P, Careen A, Edwards L, Bartlett S, McLaughlin JR, Younghusband HB. 2007. A new variant database for mismatch repair genes associated with Lynch syndrome. *Hum Mutat* 28:669-73.
- Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11:377-94.
- Young J, Barker MA, Simms LA, Walsh MD, Biden KG, Buchanan D, Buttenshaw R, Whitehall VL, Arnold S, Jackson L and others. 2005. Evidence for BRAF mutation and variable levels of microsatellite instability in a syndrome of familial colorectal cancer. *Clin Gastroenterol Hepatol* 3:254-63.
- Yuan ZQ, Wong N, Foulkes WD, Alpert L, Manganaro F, Andreutti-Zaugg C, Iggo R, Anthony K, Hsieh E, Redston M and others. 1999. A missense mutation in both hMSH2 and APC in an Ashkenazi Jewish HNPCC kindred: implications for clinical screening. *J Med Genet* 36:790-3.
- Zaanan A, Meunier K, Sangar F, Flejou JF, Praz F. 2011. Microsatellite instability in colorectal cancer: from molecular oncogenic mechanisms to clinical implications. *Cell Oncol (Dordr)*

Figure Legends

Figure 1. Colon Cancer Family Registry molecular testing schema. *somatic *BRAF* p.Val600Glu mutation testing. Note: *BRAF* results were not used in MMR gene testing selection criteria. USC: University of Southern California; QIMR: Queensland Institute of Medical Research; MSI-L/H: low/high microsatellite instability; MSS: microsatellite stable; MLPA: multiplex ligase-dependent probe amplification; DHPLC: denaturing high performance liquid chromatography.

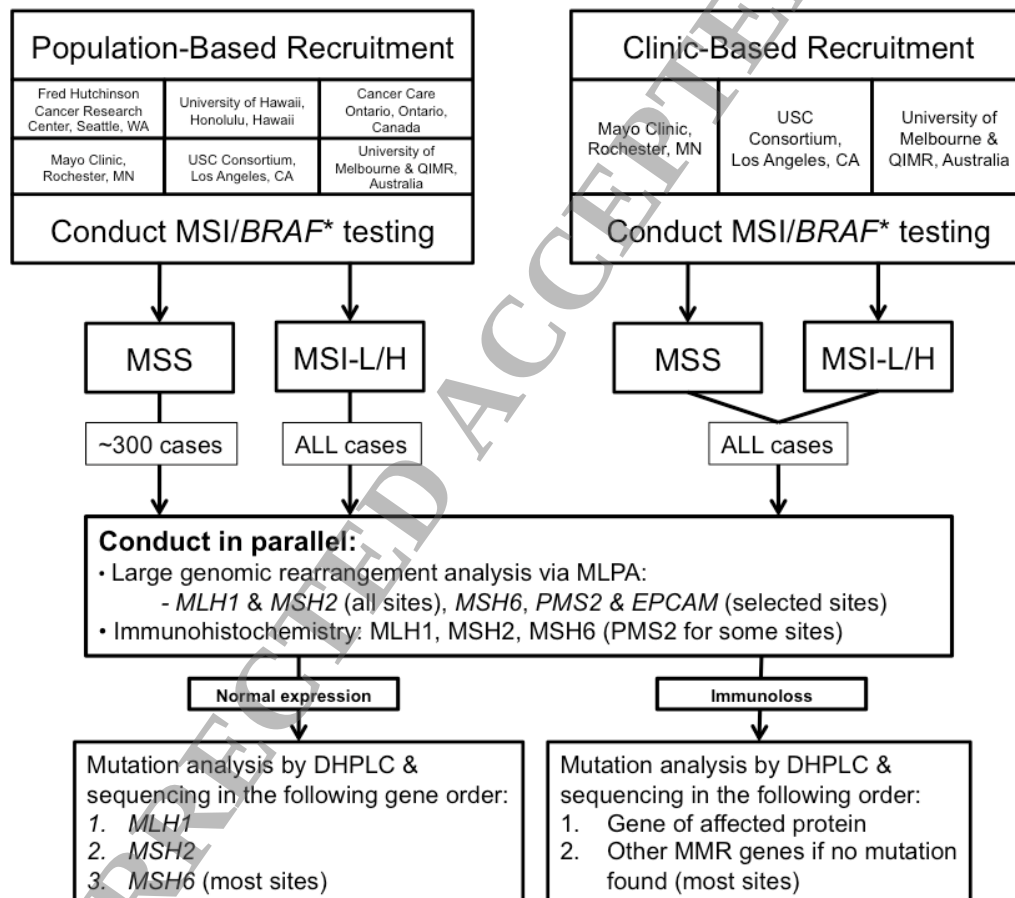


Figure 2. Splicing aberrations arising from *MLH1* and *MSH2* variants detected using RT-PCR. Variant carrier and control cycloheximide treated (+) and untreated (-) lymphoblastoid cell lines.

A) Alternative splicing detected for *MLH1* c.588+1G>T (P1), *MLH1* c.678-1G>C (P2), *MLH1* c.791-1G>C (P3), *MLH1* c.884G>A (P4) and two healthy controls (C1 and C2). The 227 base-pair insertion of *MLH1* intron 1, 145 base-pair insertion of *MLH1* intron 2, Δ exon 3, Δ exon 6, Δ exon 10, and Δ exon 9/10 are all naturally occurring isoforms. **B)** A schematic representing the splicing aberrations caused by *MLH1* c.588+1G>T (P1), *MLH1* c.791-1G>C (P3) and *MLH1* c.884G>A (P4). *MLH1* c.588+1G>T causes exon 7 skipping and both *MLH1* c.791-1G>C and *MLH1* c.884G>A lead to upregulation of Δ exon 10. **C)** *MLH1* c.1559-2A>T results in two aberrant transcripts not present in controls: Δ exon 14 and Δ exon 14/15. **D)** *MSH2* c.645+1G>A (P6 and P7 are both variant carriers) causes an upregulation of exon 3 skipping and 154 bp partial deletion of exon 3 as a minor transcript. **E)** *MLH1* c.678-1G>C causes a 2 bp deletion at the start of exon 8. **F)** *MSH2* c.1387-9T>A results in a 7 bp insertion of intron 8 caused by the creation of a *de novo* acceptor site. M: marker; bp: base-pair; fs: frameshift.

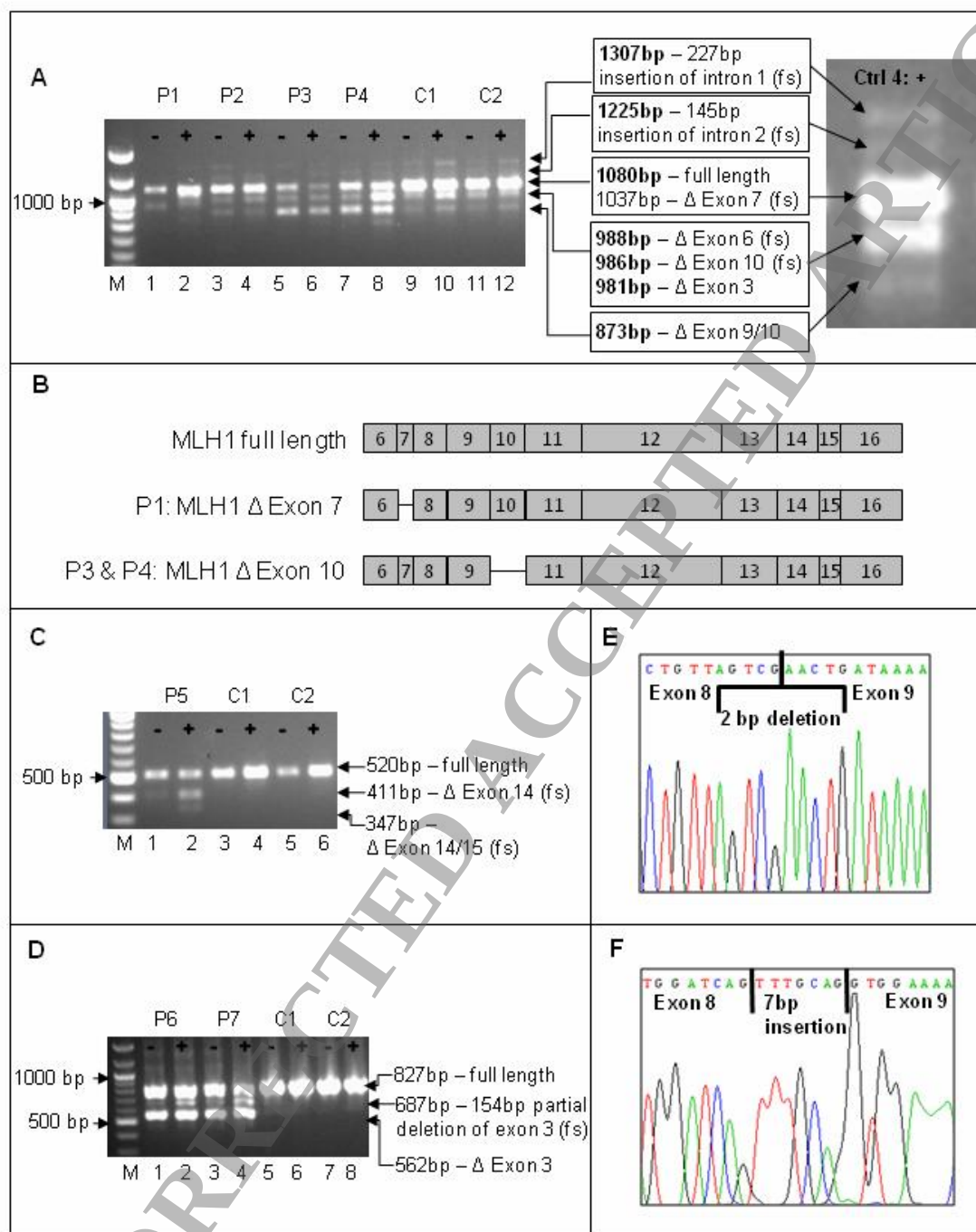


Table 1. Estimation of likelihood ratios for tumor microsatellite instability and *BRAF* mutation status

Tumor Category	Pathogenic Variant Carriers	Non-carriers	Likelihood Ratio of Causality: % MC/% NC	P-value: MC vs NC
Clinic-based	n=90	n=190		P-trend <10 ⁻⁶
MSS	3.33% (n=3)	80.00% (n=152)	0.04 (CI 95% 0.01, 0.13)	<10 ⁻⁶
MSI-L	5.56% (n=5)	9.47% (n=18)	0.59 (CI 95% 0.23, 1.5)	0.2648
MSI-H	91.11% (n=82)	10.53% (n=20)	8.66 (CI 95% 5.69, 13.2)	<10 ⁻⁶
Population-based*	n=137	n=3893		P-trend <10 ⁻⁶
MSS	7.30% (n=10)	76.03% (n=2960)	0.10 (CI 95% 0.05, 0.17)	<10 ⁻⁶
MSI-L	3.65% (n=5)	11.17% (n=435)	0.33 (CI 95% 0.14, 0.78)	0.0055
MSI-H	89.05% (n=122)	12.8% (n=498)	6.96 (CI 95% 6.29, 7.70)	<10 ⁻⁶
Clinic-based	n=69	n=14		
MSI-H <i>BRAF</i> +	2.90% (n=2)	28.57% (n=4)	0.10 (CI 95% 0.02, 0.50)	0.005
MSI-H <i>BRAF</i> -	97.10% (n=67)	71.43% (n=10)	1.36 (CI 95% 0.97, 1.90)	

Population-based*	n=52	n=189		
MSI-H <i>BRAF</i> +	3.85% (n=2)	47.62% (n=90)	0.08 (CI 95% 0.02, 0.32)	0.0007
MSI-H <i>BRAF</i> –	96.15% (n=50)	52.38% (n=99)	1.84 (CI 95% 1.59, 2.13)	

*As noted in Figure 1, tumor MSI-H and MSI-L status was used to select population-based cases for MMR sequencing, with only a subset of population-based MSS cases selected for sequencing. Therefore, the number of pathogenic variant carriers in the total population-based sample set was estimated separately for each MSI stratum based on the data available for probands that had undergone sequencing. See Methods and Supp. Table S4 for more detailed explanation. Frequency of *BRAF* mutation in MSI-H cases differs between population-based and clinic-based cases (Mantel-Haenszel X^2 : 13.83, $p=0.0002$). *BRAF* +: positive for *BRAF* p.Val600Glu mutation; *BRAF* -: negative for *BRAF* p.Val600Glu mutation; MC: mutation carrier; NC: non-carrier; CI: confidence interval.

Table 2. Mismatch repair gene multifactorial likelihood analysis results

Variant	Prior probability of pathogenicity ^f	Tumor characteristics LR	Segregation LR	Odds for causality	Posterior probability of pathogenicity	<i>In vitro</i> splicing results	Classification ^g
MLH1							
c.112A>C p.(Asn38His)	0.750	3.73x10 ⁸	1871.3	6.97x10 ¹¹	1.000	Wild-type	Class 5 - Pathogenic ¹
c.113A>G p.(Asn38Ser)	0.723	6.960	27.47	191.2	0.998	Wild-type	Class 5 - Pathogenic
c.116+5G>C	0.260	75.00	21.68	1625.9	0.998	Intron inclusion	Class 5 - Pathogenic
c.198C>T ^a	0.100	N/A	0.690	0.690	0.071	Wild-type	Class 3 - Uncertain
c.199G>A p.(Gly67Arg)	0.962 → 0.9	2210.3	2.148	4747.2	1.000	Wild-type	Class 5 - Pathogenic
c.230G>A p.(Cys77Tyr)	0.880	11.78	3.161	37.23	0.996	Not tested	Class 5 - Pathogenic
c.303T>G	0.100	N/A	0.531	0.531	0.056	Wild-type	Class 3 - Uncertain
c.307-29C>A ^a	0.260	4.10x10 ⁻¹²	4.08x10 ⁻⁴	1.67x10 ⁻¹⁵	5.88x10 ⁻¹⁶	Wild-type	Class 1 - Not pathogenic
c.350C>T p.(Thr117Met)	0.893	1.53x10 ¹¹	1.04x10 ⁶	1.60x10 ⁷	1.000	Wild-type	Class 5 - Pathogenic
c.389A>G p.(Tyr130Cys) ^b	0.757	1.360	1.1782	1.602	0.833	Wild-type	Class 3 - Uncertain
c.588+1G>T	0.960	883.3	3.706	3273.4	1.000	Exon skipping	Class 5 - Pathogenic

c.589-2A>G	0.960	N/A	25.11	25.11	0.998	4 bp deletion	Class 5 - Pathogenic
c.678-1G>C	0.960	N/A	1.671	1.671	0.976	2 bp deletion	Class 4 - Likely pathogenic
c.702G>A ^c	0.100	N/A	0.949	0.949	0.095	Wild-type	Class 3 - Uncertain
c.790+2dupT	0.960	6.960	312.8	2176.7	1.000	Exon skipping	Class 5 - Pathogenic
c.790+10A>G ^a	0.260	4.00x10 ⁻⁴	N/A	4.00x10 ⁻⁴	1.41x10 ⁻⁴	Wild-type	Class 1 - Not pathogenic ²
c.791-1G>C	0.960	8.660	122.2	1057.8	1.000	Exon skipping	Class 5 - Pathogenic
c.884G>A p.(Ser295Asn)	0.723	11.78	2.430	28.62	0.987	Exon skipping	Class 5 - Pathogenic
c.1040C>A p.(Thr347Asn)	0.007 → 0.1	N/A	1.61x10 ⁻³	1.61x10 ⁻³	1.79x10 ⁻⁴	Not tested	Class 1 - Not pathogenic
c.1321G>A p.(Ala441Thr) ^b	0.011 → 0.1	1.360	0.577	0.784	0.080	Not tested	Class 3 - Uncertain
c.1559-2A>T	0.960	102.0	1.762	179.7	1.000	Exon skipping	Class 5 - Pathogenic
c.1852_1853delAAins GC p.(Lys618Ala)	0.899	2.78x10 ⁻¹³	1.17x10 ⁻¹	3.26x10 ⁻¹⁴	2.89x10 ⁻¹³	Wild-type	Class 1 - Not pathogenic
c.1852_1854delAAG p.(Lys618del)	0.500	5624.3	579.9	3.26x10 ⁶	1.000	Wild-type	Class 5 - Pathogenic
c.1990-1G>A	0.96	N/A	28.44	28.44	0.999	Exon skipping	Class 5 - Pathogenic ³
c.2059C>T p.(Arg687Trp)	0.429	8.660	4.769	41.30	0.969	Wild-type	Class 4 - Likely pathogenic ^{3,4}
c.2146G>A p.(Val716Met) ^a	0.327	4.16x10 ⁻⁴	0.204	8.48x10 ⁻⁵	4.12x10 ⁻⁵	Wild-type	Class 1 - Not pathogenic ^{2,5,6,7,8}
c.2153A>C p.(His718Pro)	0.931 → 0.9	N/A	0.811	0.811	0.880	Wild-type	Class 3 - Uncertain

MSH2							
c.6G>T	0.100	0.100	0.696	0.070	0.008	Not tested	Class 2 – Likely Not pathogenic
c.339G>A ^a	0.100	0.004	0.119	4.75×10^{-4}	5.27×10^{-5}	Wild-type	Class 1 - Not pathogenic
c.484G>A p.(Gly162Arg) ^d	0.962 → 0.9	8.660	1.636	14.17	0.992	Wild-type	Class 5 - Pathogenic
c.488T>G p.(Val163Gly) ^e	0.422	0.087	25.56	2.213	0.618	Wild-type	Class 3 - Uncertain
c.571_573delCTC p.(Leu191del)	0.500	11.78	19.81	233.3	0.996	Wild-type	Class 5 - Pathogenic
c.645+1G>A	0.960	11.78	7.870	92.70	1.000	154 bp deletion	Class 5 - Pathogenic
c.913G>A p.(Ala305Thr)	0.929 → 0.9	N/A	0.520	0.520	0.824	Wild-type	Class 3 - Uncertain
c.942+3A>T	0.260	N/A	20871.32	20871.32	1.000	Exon skipping	Class 5 - Pathogenic
c.1013G>A p.(Gly338Glu)	0.959 → 0.9	11.78	1.350	15.90	0.993	Wild-type	Class 5 - Pathogenic
c.1387-9T>A	0.260	75.00	1.682	126.1	0.978	Intron inclusion	Class 4 – Likely pathogenic
c.1662-9G>A	0.26	2.80×10^{-12}	N/A	2.80×10^{-12}	9.85×10^{-13}	Wild-type	Class 1 - Not Pathogenic
c.1666T>C	0.1	5.86×10^{-6}	N/A	5.86×10^{-6}	6.52×10^{-7}	Wild-type	Class 1 - Not pathogenic
c.1865C>T p.(Pro622Leu)	0.927 → 0.9	75.00	67134.69	75.00	0.999	Wild-type	Class 5 - Pathogenic
c.1906G>C p.(Ala636Pro)	0.045 → 0.1	2.42×10^{-11}	4.142	1.00×10^{-2}	1.000	Wild-type	Class 5 - Pathogenic ^{9,10}
c.2005+8dupA ^a	0.260	1.840	N/A	1.840	0.393	Wild-type	Class 2 - Likely Not pathogenic
c.2006-36_2006-33dupTTAA	0.260	N/A	1.000	1.000	0.260	Not tested	Class 3 - Uncertain

c.*129T>C	0.260	0.003	0.793	0.002	7.72x10 ⁻⁴	Not tested	Class 1 - Not pathogenic
-----------	-------	-------	-------	-------	-----------------------	------------	--------------------------

Table 2. Continued

Variant	Prior probability of pathogenicity ^f	Tumor characteristics LR	Segregation LR	Odds for causality	Posterior probability of pathogenicity	<i>In vitro</i> splicing results	Classification ^g
c.*221G>T ^a	0.260	0.040	1.351	0.054	0.019	Not tested	Class 2 - Likely Not pathogenic
c.*226A>G ^a	0.260	0.080	N/A	0.080	0.027	Not tested	Class 2 - Likely Not pathogenic
MSH6							
c.1019T>C p.(Phe340Ser)	0.002 → 0.1	0.100	1.000	0.100	0.011	Wild-type	Class 2 - Likely Not pathogenic
c.1164C>T ^d	0.100	N/A	1.07x10 ⁻⁴	1.07x10 ⁻⁴	1.19x10 ⁻⁵	Wild-type	Class 1 - Not pathogenic
c.1186C>G p.(Leu396Val) ^a	0.026 → 0.1	0.100	0.520	0.052	0.006	Wild-type	Class 2 - Likely Not pathogenic
c.1193T>A p.(Val398Glu)	0.913 → 0.9	11.78	1.523	17.93	0.994	Not tested	Class 5 - Pathogenic
c.2057G>A p.(Gly686Asp)	0.916 → 0.9	1.840	1.998	3.676	0.971	Wild-type	Class 4 - Likely pathogenic
c.2341C>T p.(Pro781Ser) ^c	0.826	N/A	1.040	1.040	0.831	Wild-type	Class 3 - Uncertain
c.3226C>T p.(Arg1076Cys) ^e	0.805	N/A	1.113	1.113	0.821	Wild-type	Class 3 - Uncertain
c.*24_28delGTTGA	0.260	N/A	0.907	0.907	0.242	Wild-type	Class 3 - Uncertain

^a Co-occurrence or co-observation of variant with a pathogenic MMR gene variant in some families; ^b Co-occurrence of both variants in the proband (phase unknown); ^{c,d} Co-observation of both variants in the proband; ^e Co-observation of both variants in a single family; ^f Prior probability of pathogenicity <0.1 and >0.9 were set at 0.1 and 0.9 respectively. The prior was derived using the following equation $\text{Logit}(\text{Pr}) = -9.20 + 2.27(\text{M}) + 4.26(\text{P})$, where

Pr is the Prior probability, M is ln(MAPP score), and P is the “pph2 probability” output by PolyPhen-2.1 (retrained without MMR gene data); Tumor characteristics LR based on colorectal tumors only; The Odds for Causality are calculated as the product of the individual statistically independent components (Tumor Characteristics LR X Segregation LR). The Posterior Probability = Posterior Odds/(Posterior Odds + 1), where the Posterior Odds = Prior Probability X Odds for causality X (1/1-prior probability);

^g Based on the criteria in Plon et al., 2008, additional data was extracted from peer-reviewed literature to achieve classification for six variants: 1 - Van Riel et al., 2010; 2 - Pastrello et al., 2011; 3 - Caldes et al., 2002; 4 - Gallinger et al., 2004; 5 - Genuardi et al., 1998; 6 - Muller-Koch et al., 2001; 7 - Kamory et al., 2006; 8 - Barnetson et al., 2008; 9 - Yuan et al., 1999; 10 - Foulkes et al., 2002;

N/A - Not applicable, LR - likelihood ratio

UNCORRECTED ACCEPTED ARTICLE