ANALYSIS

nature genetics

Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database

Bryony A Thompson[1,2,46], Amanda B Spurdle[1,46], John-Paul Plazzer[3], Marc S Greenblatt[4], Kiwamu Akagi[5], Fahd Al-Mulla[6], Bharati Bapat[7], Inge Bernstein[8,9], Gabriel Capellá[10], Johan T den Dunnen[11], Desiree du Sart[12], Aurelie Fabre[13], Michael P Farrell[14], Susan M Farrington[15], Ian M Frayling[16], Thierry Frebourg[17,18], David E Goldgar[19,20], Christopher D Heinen[21,22], Elke Holinski-Feder[23,24], Maija Kohonen-Corish[25–27], Kristina Lagerstedt Robinson[28], Suet Yi Leung[29], Alexandra Martins[17], Pal Moller[30], Monika Morak[23,24], Minna Nystrom[31], Paivi Peltomaki[32], Marta Pineda[10], Ming Qi[33,34], Rajkumar Ramesar[35], Lene Juel Rasmussen[36], Brigitte Royer-Pokora[37], Rodney J Scott[38,39], Rolf Sijmons[40], Sean V Tavtigian[20], Carli M Tops[11], Thomas Weber[41], Juul Wijnen[11], Michael O Woods[42], Finlay Macrae[3] & Maurizio Genuardi[43,44] on behalf of InSiGHT[45]

The clinical classification of hereditary sequence variants identified in disease-related genes directly affects clinical management of patients and their relatives. The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) undertook a collaborative effort to develop, test and apply a standardized classification scheme to constitutional variants in the Lynch syndrome–associated genes *MLH1*, *MSH2*, *MSH6* and *PMS2*. Unpublished data submission was encouraged to assist in variant classification and was recognized through microattribution. The scheme was refined by multidisciplinary expert committee review of the clinical and functional data available for variants, applied to 2,360 sequence alterations, and disseminated online. Assessment using validated criteria altered classifications for 66% of 12,006 database entries. Clinical recommendations based on transparent evaluation are now possible for 1,370 variants that were not obviously protein truncating from nomenclature. This large-scale endeavor will facilitate the consistent management of families suspected to have Lynch syndrome and demonstrates the value of multidisciplinary collaboration in the curation and classification of variants in public locus-specific databases.

Identification of a high-risk disease-causing constitutional mutation in a cancer patient guides the clinical management of their whole family, with implications for counseling, cancer treatment options and presymptomatic surveillance, and considerations of risk-reducing surgery and/or medication regimens[1]. Carriers of mutations in the mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* and *PMS2* causing Lynch syndrome[1] have a substantially increased risk of colorectal and endometrial cancers, along with increased risk of ovarian, gastric, small bowel, urothelial, brain, hepatobiliary, pancreatic, bladder, kidney, prostate and breast cancers[1–8]. However, intensive management reduces mortality[9].

Sequence variants of uncertain functional and clinical relevance are common in genetic test reports. Although several lines of evidence can be evaluated to assess the clinical implications of these variants, usually none of these approaches can be used on its own to obtain clinically useful interpretations and, for many variants, comprehensive data are lacking. Laboratories are generally conservative in designating pathogenic variants, defining variants as being of 'uncertain significance' unless overwhelming evidence of pathogenicity exists. Several schemes for the classification of variants in genes associated with mendelian conditions have been proposed for use in the clinical setting. Because clinically useful actions are currently only considered for high-penetrance mutations, all of these systems are aimed at differentiating high-penetrance from low-penetrance and neutral variants and do not consider variants of intermediate risk. These schemes differ in the range and format of data used for classification and in the number of variant classes[10–12]. The International Agency for Research on Cancer (IARC) classification system, endorsed by the Human Variome Project (HVP), facilitates standardized categorization by defining classes that can be linked to validated quantitative measures of causality and/or pathogenicity from statistical models[13–16] or to validated interpretation of qualitative data[17]. Importantly, only the five-class IARC system has been linked to clinical recommendations for all classes, including clinical testing and full high-risk surveillance guidelines for class 5 (pathogenic) and class 4 (likely pathogenic) variants; advice to treat as if "no mutation associated with disease has been detected" for class 1 (not pathogenic) and class 2 (likely not pathogenic) variants; and

A full list of authors and affiliations appears at the end of the paper.

Received 19 September; accepted 26 November; published online 22 December 2013; doi:10.1038/ng.2854

acquisition of additional data to provide more robust classification for class 2 (likely not pathogenic), class 3 (uncertain) and class 4 (likely pathogenic) variants.

Locus-specific databases (LSDBs) are an important source of information for clinicians and researchers in assessing data and forming opinions on the clinical relevance of disease-associated gene sequence variants, and these databases have a fundamental role in variant classification owing to their added value from having aggregated data. Consistent and normalized data curation is critical to the value derived from databases in categorizing the relationship between genetic variation and disease—especially for clinical applications. It has previously been recommended by the IARC Working Group that a panel covering a range of expertise in variant classification provide consensus opinion on variant pathogenicity before publicly accessible display of such information[18]. Another important component of the classifications provided by LSDBs is transparency regarding the criteria and supporting information used for classification, which allows LSDB users to consider the information for their own applications in research or clinical settings[18].

InSiGHT has merged multiple gene mutation and variant repositories to create the InSiGHT Colon Cancer Gene Variant Database for MMR and other colon cancer susceptibility genes[19–23], hosted by the Leiden Open Variation Database (LOVD). Following recommendations for LSDB curation[18], InSiGHT formed an international panel of researchers and clinicians to review MMR gene variants submitted to the database. To encourage the submission of unpublished clinical and research data to further facilitate variant classification, the microattribution approach[24] was implemented using Open Researcher and Contributor Identification (ORCID). Here we present the results of the InSiGHT Variant Interpretation Committee (VIC) effort to develop, test and apply a 5-tiered scheme to classify 2,360 unique constitutional MMR gene variants.

RESULTS

Curation of MMR gene variants

As of the end of December 2012, after 3,458 alterations to standardize nomenclature had been made, there were 12,635 submissions of 2,730 unique MMR gene variants lodged in the InSiGHT database. Furthermore, 370 unique variants (13.6%) were not identified in constitutional (germline) DNA (see **Supplementary Fig. 1** and **Supplementary Table 1** for details) and were excluded from further analyses because (i) no evidence existed that these occurred as constitutional variants and (ii) no clinical information was available to assess their potential roles in hereditary disease. The 2,360 constitutional variants included 932 *MLH1* (39%), 842 *MSH2* (36%), 449 *MSH6* (19%) and 137 *PMS2* (6%) variants. Most variants (800; 34%) were nonsense or frameshift changes predicted to cause protein truncation, with the next largest group (746; 32%) consisting of nonsynonymous variants that were not obviously truncating, including missense substitutions, small in-frame insertion-deletion mutations (indels) and read-through alterations of the translation termination codon.

Variants had originally been assigned a classification by submitters according to the following classes: pathogenic, probably pathogenic, no known pathogenicity, probably no pathogenicity or effect unknown. No information was recorded to document the rationale for classification or the standards used to classify variants. When considering the 1,382 constitutional variants with multiple entries in the InSiGHT database, discordance in classification between submitters was found for 869 variants. Some of this discordance arose because of classification based on single data points or references, such as results from a single functional assay[22] or inferences from individual publications originally lodged in the Mismatch Repair Genes Variant Database[23] (see the example in **Supplementary Table 2**).

Table 1 InSiGHT variant classification scheme with accompanying recommendations for family management, adapted from the IARC five-tiered classification system

InSiGHT MMR gene variant class definition for Lynch syndrome[a]	Predictive testing of at-risk relatives	Surveillance for at-risk relatives	Research testing of relatives
5 (pathogenic)	Yes	Full high-risk guidelines	Not indicated
4 (likely pathogenic)	Yes[b]	Full high-risk guidelines	Yes
3 (uncertain)	No[b]	Family history and other risk factors	Yes
2 (likely not pathogenic)	No[b]	Family history and other risk factors; treated as having no mutation detected in this gene for this disorder	Yes
1 (not pathogenic)	No[b]	Family history and other risk factors; treated as having no mutation detected in this gene for this disorder	Not indicated

Adapted from Plon *et al.*[10]. Full high-risk surveillance guidelines for cancers in the Lynch syndrome spectrum are outlined in Vasen *et al.*[1]. Research testing entails cascade testing for the variant in affected and unaffected family members to facilitate segregation analysis and is indicated for variants in classes 2–4 to refine classification. Consent from subjects through a protocol approved by a human subjects committee should be obtained.

[a]Class definition is described in detail in **Supplementary Table 4**, and the **Supplementary Note** and is based on quantitative evidence defined by multifactorial likelihood posterior probability (with cutoffs of >0.99 for class 5, 0.95–0.99 for class 4, 0.05–0.949 for class 3, 0.001–0.049 for class 2 and <0.001 for class 1) or combined qualitative evidence determined by consensus opinion, as defined by the InSiGHT VIC. Pathogenic variants are defined as being clinically relevant in a genetic counseling setting, such that germline variant status will be used to inform patient and family management. [b]Continued testing of the proband is recommended for any additional available testing modalities available, for example, rearrangements (Online Methods).

Development of a five-tiered system for classification

The InSiGHT VIC (Online Methods) was established in 2007 to address discrepancies in the classification of MMR gene variants lodged in the InSiGHT database. Since March 2011, the VIC has made a concerted effort to develop standardized criteria for variant classification, employing a modified Delphi consensus process[25] to evaluate current scientific evidence and reach consensus. In line with HVP[26], the IARC classification system[10] for variant categorization (**Table 1**) was adopted by InSiGHT for the classification of MMR gene variants. Briefly, multiple lines of evidence were required for classification, and evidence for each variant had to include data associating the variant with both clinical and functional consequences (Online Methods).

The scheme was first tested on a subset of 117 MMR gene variants, and the criteria evolved and were refined by consensus to accommodate new data and inconsistencies over multiple classification teleconferences and face-to-face meetings. Final criteria were then applied retrospectively and to all remaining unique variants listed in the database (**Supplementary Table 3**). An overview of the InSiGHT classification criteria is shown in **Figure 1** (see **Supplementary Table 4** and the **Supplementary Note** for detailed criteria and justifications). At the close of each VIC teleconference or meeting, consensus classifications were noted. Where necessary, action items to improve or clarify classification included (i) calls for missing clinical and functional information for specific variants to committee members and the general InSiGHT membership; (ii) requests for more detailed data or data clarification from the authors of original publications; and (iii) reassessment of classification after additional data were obtained. At the end of the process, the InSiGHT database was updated with the

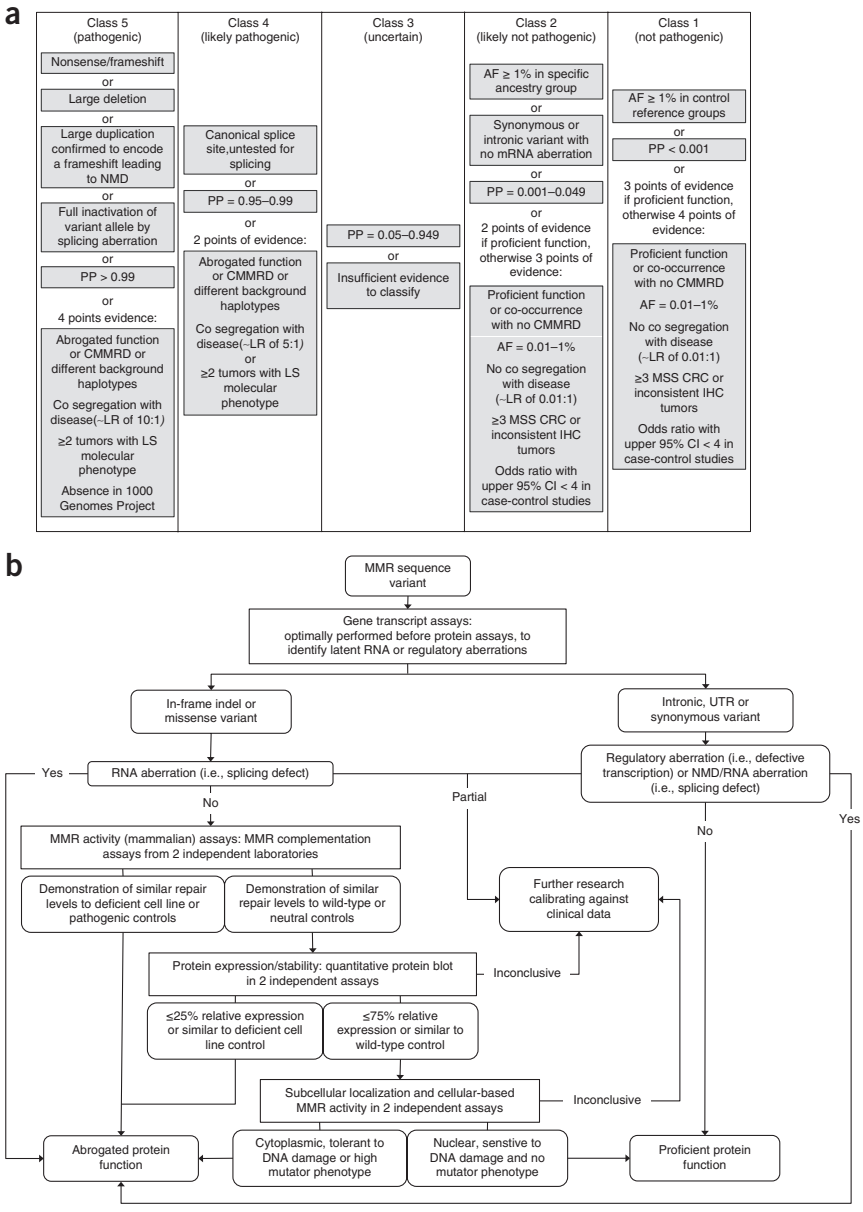


Figure 1 Overview of the five-tiered InSiGHT classification guidelines. (**a**) Simplified guidelines describing the levels and types of evidence required for each tier of the classification. Full guidelines are described in the **Supplementary Note**, and the detailed rationale behind each criterion is provided in **Supplementary Table 3**. The Lynch syndrome molecular phenotype described in classes 4 and 5 includes information on MSI and/or loss of expression of relevant protein(s), as determined by immunohistochemistry. In this study, variants resulting in the introduction of a premature termination codon or large genomic deletions affecting functionally important domains, generally considered pathogenic on the basis of DNA sequence alone, are referred to as class 5a (assumed pathogenic) variants. All other variants categorized in class 5 are termed class 5b variants. (**b**) Flowchart used to assist in the interpretation of available data from functional assays. Assays reviewed for classification are shown in **Supplementary Table 4**, and the values used to define abrogated or normal function are shown in **Supplementary Table 5**. Cutoffs of <25% and >75% protein expression, as used in previous publications[47,48], are very conservative given that abrogated function has been reported to be associated with a decrease in MLH1 expression of ~50% or more[49]. For variants that had normal, inconclusive or intermediate MMR activity in two independent assays but were deficient in protein function in two independent assays, abrogated function was assigned. AF, allele frequency; PP, posterior probability of pathogenicity derived by multifactorial likelihood analysis; CMMRD, constitutional MMR deficiency (MIM 276300); LR, likelihood ratio; LS, Lynch syndrome; MSS, microsatellite stable; CRC, colorectal cancer; IHC, immunohistochemistry; NMD, nonsense-mediated decay.

final consensus classifications and the supporting data to ensure transparency.

The major issues faced by the committee in the review process included redundancy across multiple sources (resolved through discussion with the authors of original publications), paucity of information, incomplete or inaccurate data and difficulties in the interpretation of the results of functional assays. To facilitate the interpretation of findings from functional assays, supporting information and flowcharts were developed (**Fig. 1b** and **Supplementary Tables 5** and **6**), and multiple meetings were coordinated that were dedicated to the review of variants with apparently discordant results from functional assays (**Supplementary Table 3**).

Validation of the InSiGHT qualitative classification criteria

Nonsense or frameshift alterations or large genomic deletions interrupting functionally important domains are generally considered pathogenic on the basis of DNA sequence alone; these variants were designated here as class 5a (assumed pathogenic). There were 990 assumed pathogenic variants in the database, 640 of which were private mutations. To demonstrate the robustness of the qualitative classification criteria, 170 assumed pathogenic variants (68 *MLH1*, 75 *MSH2*, 13 *MSH6* and 14 *PMS2*) were reviewed as a validation set against the class 5 (pathogenic) qualitative criteria required for the assignment of variants to class 5b (Online Methods and **Supplementary Table 7**). Designation of a variant as class 5b required evidence of abrogated protein function, at least two tumors with microsatellite instability (MSI) or appropriate loss of MMR protein expression and a segregation likelihood ratio of >10:1 (incorporating gene-specific cumulative risks[27]) or variant cosegregation with disease reported in at least two Amsterdam criteria–positive families. Class 5b was attained by all 60 validation set variants that had sufficient clinical data to assess these required criteria. The other 110 validation set variants could not be assigned to class 5b, largely because family cosegregation and tumor data were scarce or unobtainable—presumably because these variants are accepted as disease causing and are routinely used for clinical presymptomatic testing in families (Online Methods). Of these variants, 72 were assigned to class 4 owing to lack of only 1 point of evidence, and 38 variants were assigned to class 3 owing to insufficient data. However, only 2 of 13 *MSH6* and 2 of 14 *PMS2* variants fulfilled class 5b criteria, reflecting the lower penetrance and later age of onset associated with deleterious variants in *PMS2* (ref. 28)

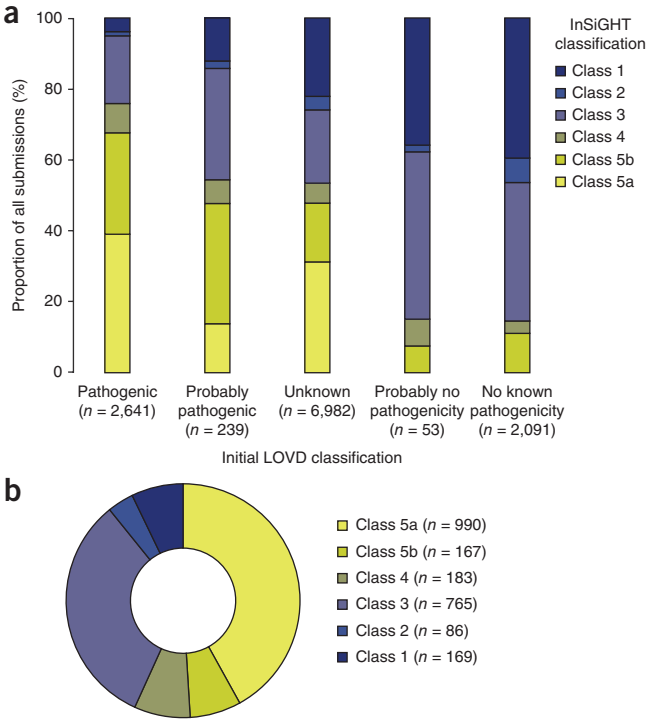


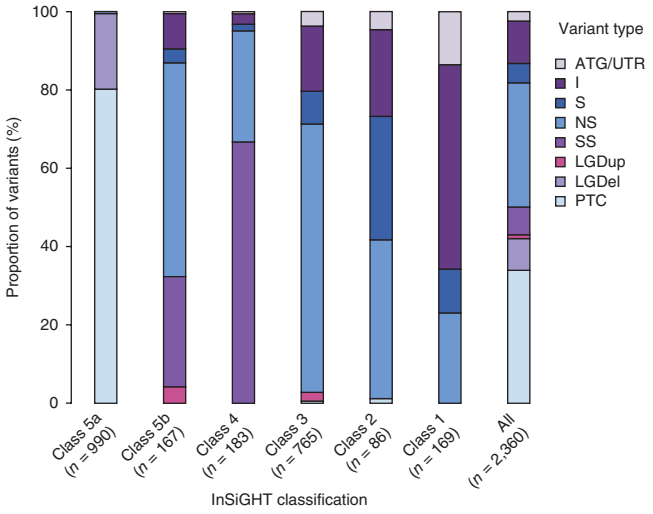
Figure 2 Outcome of standardized five-tiered InSiGHT classification of constitutional MMR gene variants. (**a**) The graph plots the proportion of five-tiered classifications for all documented constitutional variants in the database against the original LOVD database classifications assigned by submitters. Class 5a is a subset of class 5 containing assumed pathogenic nonsense mutations, small frameshift indels and large deletions. Class 5b includes variants that are not obviously truncating but are considered to be pathogenic on the basis of combined evidence (**Supplementary Note**). Results show that standardized classification led to altered classifications for a considerable proportion of variant entries, including the downgrading of variants submitted as pathogenic (24%) and the upgrading of variants with unknown pathogenicity to likely pathogenic (5.6%) or pathogenic (48%). In addition, clinically important misclassifications were identified for unique variants initially submitted as not pathogenic (54 unique variants reclassified as class 5b variants and 25 reclassified as class 4 variants) and unique variants submitted as pathogenic (28 unique variants reclassified as class 1 variants, 16 reclassified as class 2 variants and 218 reclassified as class 3 variants). (**b**) Pie chart showing the distribution of final InSiGHT VIC consensus classifications.

and *MSH6* (ref. 29). Together, these results indicate that the criteria for classification using qualitative data were sufficiently stringent to ensure conservative classification.

Classification of 2,360 constitutional MMR gene variants

Of the 12,006 eligible variant entries in the InSIGHT database, submitter and final classifications differed for 7,935 (66%), including changes from class 1 (not pathogenic) to class 5 (pathogenic) and vice versa (**Fig. 2a**). The overall breakdown of final classifications is shown in **Figure 2b**. In addition to the 990 assumed pathogenic truncating or large-deletion variants (class 5a), consistent medical management is now also possible for the remaining 1,370 not obviously truncating variants; these include 167 class 5 (pathogenic) variants (class 5b) (12%), 183 class 4 (likely pathogenic) variants (14%), 86 class 2 (likely not pathogenic) variants (6%) and 169 class 1 (not pathogenic) variants (12%).

Nonsynonymous variants made up the majority of class 3 variants (524/765; 68%) and newly assigned class 5b variants (91/167; 54%) (**Fig. 3** and **Supplementary Fig. 2**; see **Supplementary Table 8** for detailed information supporting classifications). Substitutions at canonical dinucleotide splice sites fell predominantly in class 4 owing to lack of functional RNA analyses; however, if experimentally tested, these variants would likely be moved to class 5b. Intronic variants outside of conserved splice sites were the most prevalent variant type in class 1.



Final categorization (Online Methods) of not obviously truncating variants as class 1, 2, 4 or 5 was achieved by applying qualitative criteria for 391 variants, quantitative multifactorial likelihood analysis methodology (based on bioinformatic prior probabilities and evidence from segregation and/or tumor data; see Thompson *et al.*[16]) for 192 variants and either quantitative or qualitative criteria for 26 variants. Where classifications derived using quantitative criteria differed from those generated with qualitative criteria, this difference reflected the amount of data available rather than deficiencies in the classification criteria, with no variants considered to fall into class 1 or 2 using one approach and class 4 or 5 using the other. Six synonymous variants were categorized in class 5b owing to their effects on splicing. Of the substitutions occurring in initiation codons (often assumed to be pathogenic[30–32]), only one of nine had sufficient evidence to determine pathogenicity.

Implementing microattribution

Microattribution is a means to incentivize the placement of unpublished data in the public domain by assigning scholarly contribution to authors similar to the citation conventions afforded to journal

Figure 3 Classifications of all documented unique variants by variant type. The plot represents the proportion of the different variant types within the five classes. Class 5a is a subset of class 5 containing assumed pathogenic mutations (nonsense mutations, small frameshift indels and large deletions). All other variants placed in class 5 are termed class 5b variants (**Supplementary Note**). The different variant types include the following: ATG/UTR, variants in the initiation codon and the 5′ or 3′ UTR; I, intronic variants outside of the canonical splice-site dinucleotides; S, synonymous variants; NS, not obviously truncating nonsynonymous variants outside of the Kozak consensus sequence, i.e., missense, small in-frame indel and read-through translation termination codon alterations; SS, variants in the canonical splice-site dinucleotides; LGDup, large genomic duplications; LGDel, large genomic deletions or disrupting inversions; PTC, variants that introduce premature termination codons, i.e., nonsense mutations and small frameshift indels. See **Supplementary Figure 2** for further details of variant types by MMR gene.

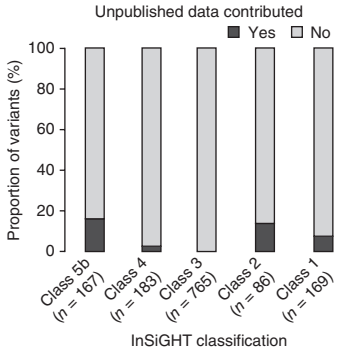


Figure 4 Contribution of microattribution to the classification of variants that are not obviously truncating. Dark shading (yes) indicates the proportion of total variants for each class where the additional data obtained through microattribution contributed to final classification.

articles[33]. Retrospective and prospective microattribution was implemented to acknowledge and encourage the submission of unpublished data to the InSiGHT database, including the submission of additional detailed clinical information from authors of published reports. Microattribution was assigned for initial variant submission, segregation and family history data, pathology diagnostics (MSI analysis and immunohistochemistry) information, data from *in vitro* functional assays (mainly RNA splicing) and the determination of variant frequencies in normal individuals. As of July 2013, a total of 6,015 microattributions were conferred, including 3,763 for variant submission, 2,111 for family and tumor pathology data, 97 for data from *in vitro* assays and 25 for frequency data. Notably, 19% of the microattributions for clinical and functional data contributed additional information critical for the classification of variants in the class 5a (assumed pathogenic) validation subset. These data also highlighted the fact that clinical testing for these assumed pathogenic variants is mostly undertaken in the presymptomatic setting. The contribution of microattribution to the final classification of not obviously truncating variants is shown in **Figure 4**. Notably, classification was altered for 57 of 169 (34%) variants for which unpublished data were obtained. Moreover, implementation of microattribution stimulated the submission of 128 new MMR gene variants yet to be classified.

Class 3 variants of uncertain relevance

Missense variants in MMR genes are abundant among class 3 (uncertain) variants and present a considerable clinical problem. Quantitative multifactorial likelihood analysis is an effective approach for the classification of missense variants, as validated bioinformatic predictions[34] based on amino acid conservation and physicochemical properties can be used as a surrogate for the *in vitro* effects of variants on protein function. *In silico* analyses previously shown to be highly accurate (area under receiving operator characteristic (ROC) curve of 0.93)[34] were used to estimate the prior probability of pathogenicity for all 481 class 3 (uncertain) missense variants (**Fig. 5**) to prioritize requests for data to facilitate future multifactorial analysis. The distribution of prior probabilities for *MLH1* and *MSH2* class 3 variants was clearly bimodal, suggesting that ~50% of *MLH1* and *MSH2* missense variants may be classified as pathogenic after further investigation. In total, 401 missense variants had extreme prior probabilities of ≤20% or ≥80%, with 270 having probabilities of <10% or >90%, indicating that classification of a variant as class 1 or class 5 could be easily performed by incorporating segregation or tumor information. It is also possible that some class 3 variants with low prior probability of pathogenicity based on the predicted missense alteration would cause splicing aberrations, as already observed for 42 of the 746 not obviously truncating nonsynonymous variants. Incorporation of validated bioinformatic splicing prediction tools into the MMR gene multifactorial model, as is under development for *BRCA1* and *BRCA2* (ref. 35), will assist in the prioritization of variants with the potential to disrupt splicing.

In investigation of the potential effects of class 3 regulatory variants (Online Methods), all 15 5′ UTR variants fell within multiple transcription factor binding sites, but no evidence for interruption of microRNA binding was found for 6 variants in the 3′ UTR (data not shown). Multifactorial analyses and transcription assays would help elucidate whether these variants affect gene function.

DISCUSSION

The InSiGHT VIC has successfully undertaken a collaborative effort to establish standardized variant interpretation guidelines, encourage data submission and provide objective assessment of MMR gene variants involved in Lynch syndrome. The criteria developed provide a basis for the standardized clinical classification of variants to inform patient and family management through genetic counseling[10]. This initiative has achieved the systematic evaluation of 2,360 constitutional variants, which will benefit thousands of families internationally. Notably, 605 variants not resulting in the introduction of a premature termination codon, including 217 nonsynonymous substitutions, have now been assigned to class 5 (pathogenic) and class 4 (likely pathogenic) or to class 1 (not pathogenic) and class 2 (likely not pathogenic). The assigned classes of these variants can now also be used as standards for the calibration of functional assays[36,37].

The clinical relevance of 32% of the variants investigated remains uncertain. A large proportion of these (71%) were private variants occurring in only one family, and these variants are difficult to classify

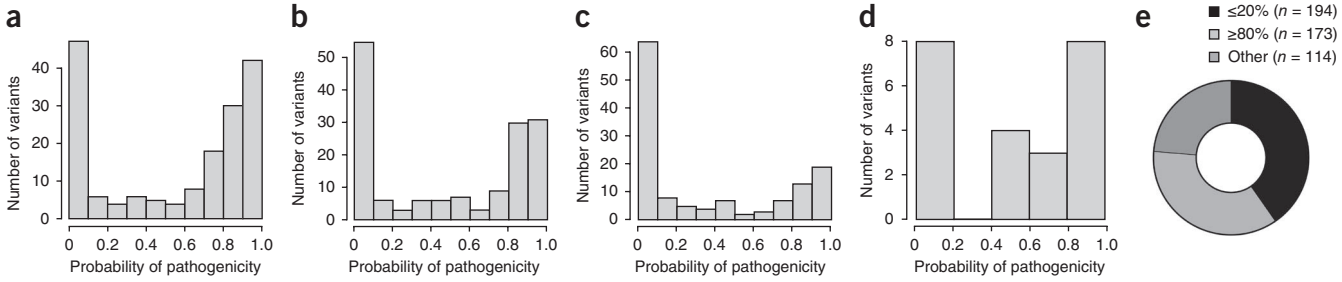


Figure 5 Probabilities of pathogenicity for 481 class 3 missense variants of uncertain effect derived by *in silico* analysis. (**a**–**e**) Distribution of probabilities of pathogenicity as estimated from a calibrated algorithm based on customized MAPP and PolyPhen-2 scores[34] for *MLH1*, $n = 186$ (**a**), *MSH2*, $n = 169$ (**b**), *MSH6*, $n = 145$ (**c**), *PMS2*, $n = 24$ (**d**) and all four MMR genes (**e**), showing stratification of variants with prior probabilities of ≤20% or ≥80% to prioritize variants for further investigation and classification.

owing to the paucity of available clinical information. Clinicians have a fundamental role in promoting the collection of segregation information and other data relevant for classification. We anticipate that the development of this interpretation scheme, along with the implementation of microattribution, will incentivize assistance in the accumulation of clinical data. The value of microattribution for data accrual has previously been demonstrated for hemoglobinopathies[24], and the InSiGHT initiative now demonstrates the clinical usefulness of data collection. The promotion of standardized data formats will assist in the transition to fully quantitative, unbiased classification, eventually incorporating other components of the qualitative guidelines. In addition, the difficulties experienced in interpreting apparently discordant data from functional assays emphasize the importance of assay validation and standardization[38,39]. Such experience will be directly applicable to the functional analysis of deep intronic and regulatory variants, which are increasingly detected with the advancement of DNA sequencing technologies.

To accommodate the lower penetrance and reported lower degrees of tumor MSI associated with *MSH6* and *PMS2* mutations[28,29,40–44], gene-specific criteria should also be considered for future iterations of the classification guidelines, for example, stipulating the inclusion of segregation odds for *MSH6* and *PMS2* variants for classification or using modified panels to detect MSI status.

Another challenging issue to contemplate will be incorporating alleles of intermediate risk[45] into classification schemes, including the clinical recommendations that might be linked to such variants. The identification of a subset of MMR gene alleles with apparently discordant clinical and functional features that renders them resistant to classification will provide the basis for future studies to define the most appropriate methodology and criteria to identify such variants. Further studies will also be required to assess whether variants resulting in abrogated DNA damage response but normal MMR[46] are associated with the same clinical features as classical pathogenic mutations in MMR genes.

The InSiGHT database is a well-recognized resource for the clinical and research communities, receiving over 20,000 hits per month. The development and adoption of standard templates allows transparency in the review process. Database users can view relevant information and sources in relation to information on guideline interpretation when considering the classification provided by the committee. The guidelines must evolve to accommodate additional kinds of evidence, but we anticipate no clinical issues as long as variant classifications are dated and linked to a dated set of guidelines with the supporting information used to derive classifications. Final classifications have also been submitted to the NCBI ClinVar database for higher exposure, but expert classifications and underlying data rest with InSiGHT.

To our knowledge, this is the first large-scale comprehensive classification effort demonstrating the value of evaluation by expert panel in the curation of an LSDB and providing summary information used to assign variant pathogenicity. This initiative also shows how classification may be assisted by promoting standardized data submission from stakeholders in the clinical and research settings, thereby allowing access to unpublished clinical and functional information used to facilitate variant classification. Thus, the InSiGHT initiative provides a key model of LSDB-centric multidisciplinary collaboration for the transparent interpretation of DNA variants.

URLs. Clinical Molecular Genetics Society (CMGS) classification system, http://cmgsweb.shared.hosting.zen.co.uk/BPGs/Best_Practice_Guidelines.htm; Human Variome Project (HVP), http://www.humanvariomeproject.org/; Leiden Open Variation Database (LOVD), http://www.lovd.nl/3.0/home; Open Researcher and Contributor Identification (ORCID), http://orcid.org/; International Society for Gastrointestinal Hereditary Tumours (InSiGHT), http://www.insight-group.org/variants/classifications/; NCBI ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/; Mutalyzer, https://mutalyzer.nl/; Huntsman Cancer Institute LOVD for MMR gene missense substitution prior probabilities of pathogenicity, http://hci-lovd.hci.utah.edu/; UCSC Genome Browser, http://genome.ucsc.edu/; Nanopub, http://www.nanopub.org/; R project, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All data can be accessed at the InSiGHT website. Variants have been submitted to LOVD and ClinVar and are searchable by the gene names *MLH1*, *MSH2*, *MSH6* and *PMS2*. The RefSeq and ClinVar accessions (respectively) for the MMR genes are as follows: *MLH1*, NM_000249.3 and NG_007109.2; *MSH2*, NM_000251.2 and NG_007110.2; *MSH6*, NM_000179.2 and NG_007111.1; *PMS2*, NM_000535.5 and NG_008466.1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all submitters of data to the InSiGHT database (retrospective and prospective), the Colon Cancer Family Registry and the German Hereditary Non-polyposis Colorectal Cancer Consortium for their contributions of unpublished data, acknowledged formally through microattribution. We would also like to acknowledge L. Marquart for providing statistical advice and T. O'Mara for providing advice and assistance with the statistical package R. We are extremely grateful to the Hicks Foundation (Australia) for inaugural support of InSiGHT database curator J.-P.P. Funding for VIC teleconferences was provided by the Cancer Council of Victoria. B.A.T. is supported by a Cancer Council of Queensland PhD scholarship and by a Queensland Institute of Medical Research PhD Top-Up award. A.B.S. is a National Health and Medical Research Council Senior Research Fellow. The work performed by A.B.S. and B.A.T. was additionally supported by Cancer Australia (1010859). M.G. is supported by a grant from the Tuscan Tumor Institute (ITT). J.-P.P. is currently supported by the Royal Melbourne Hospital Foundation. S.V.T., M.S.G., A.B.S., L.J.R. and R.S. are supported by grant 1R01CA164944 from the National Cancer Institute/US National Institutes of Health (NCI/US NIH). G.C. and M.P. were supported by the Ministerio de Ciencia e Innovación (SAF 12-33636) and by the Fundación Científica de la Asociación Española Contra el Cáncer. A.F. is supported by the French National Cancer Institute and by the Institut National du Cancer (INCa) French MMR Committee. S.M.F. is supported by grants from the Association of International Cancer Research (12-1087) and by the Medical Research Council UK (MR/K018647/1). NHS Wales National Institute for Health and Social Care (NIHSCR) funding was provided to I.M.F. via the Cardiff & Vale University Health Board. D.E.G. is supported by funding from Mayo Specialized Programs of Research Excellence (SPORE) grant P50CA11620106 (principal investigator J. Ingle). C.D.H. is funded by US NIH grant R01 CA115783-02/CA/NCI. E.H.-F. and M.M. are supported by German Cancer Aid (Deutsche Krebshilfe) and by the Wilhelm Sander Foundation. M.K.-C. is funded by Cancer Institute NSW. S.Y.L. is supported by the Hong Kong Cancer Fund. A.M. is supported by the French National Cancer Institute and by the Direction Générale de l'Offre des Soins (INCa/DGOS). The Sigrid Juselius Foundation funds M.N. Funding for P.P. is provided by the European Research Council (FP7-ERC-232635). L.J.R. is funded by Nordea-Fonden. B.R.-P. is supported by German Cancer Aid. M.O.W. was supported by the Canadian Cancer Society Research Institute (grant 18223).

AUTHOR CONTRIBUTIONS

A.B.S. and B.A.T. drafted the manuscript. B.A.T. conducted InSiGHT database nomenclature standardization and data cleaning, systematic literature and data review, statistical analyses and final data analyses and assisted in the presentation of data in web-based format. B.A.T., A.B.S., S.V.T., M.S.G., D.E.G. and M.G. formulated the baseline guidelines for consideration by VIC members. B.A.T. and A.B.S. developed the functional flowchart and, with L.J.R., C.D.H., G.C., M.P., A.M., B.R.-P., E.H.-F., M.S.G., M.M., T.F. and M.N. formed the functional

subcommittee contributing to the supporting documents for functional assay interpretation. D.E.G. provided statistical input. J.-P.P. provided data management, organized teleconferences, collated information after teleconferences, coordinated microattribution and was responsible for the presentation of data in web-based format. J.T.d.D. provided support for the LOVD database and created the LOVD nanopublications. F.M. is the responsible InSiGHT Councilor who initiated the concept of VIC in 2007 and has been responsible for advocating for funding and organizing the face-to-face meeting in Paris. M.G. coordinated VIC and chaired teleconferences and face-to-face meetings. B.A.T., A.B.S., S.V.T., M.S.G., D.E.G., M.G., F.M., L.J.R., C.D.H., G.C., M.P., A.M., B.R.-P., E.H.-F., M.S.G., M.M., T.F., M.N., K.A., F.A.-M., B.B., I.B., D.d.S., A.F., M.P.F., S.M.F., I.M.F., M.K.-C., K.L.R., S.Y.L., P.M., P.P., M.Q., R.R., R.J.S., R.S., C.M.T., T.W., J.W. and M.O.W. provided critique of the classification criteria and/or participated in review of variants at teleconferences or face-to-face meetings or by e-mail and provided critical review of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Vasen, H.F. *et al.* Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut* **62**, 812–823 (2013).
2. Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**, 261–268 (2004).
3. van Oers, J.M. *et al.* PMS2 endonuclease activity has distinct biological functions and is essential for genome maintenance. *Proc. Natl. Acad. Sci. USA* **107**, 13384–13389 (2010).
4. Win, A.K. *et al.* Risks of primary extracolonic cancers following colorectal cancer in lynch syndrome. *J. Natl. Cancer Inst.* **104**, 1363–1372 (2012).
5. Buerki, N. *et al.* Evidence for breast cancer as an integral part of Lynch syndrome. *Genes Chromosom. Cancer* **51**, 83–91 (2012).
6. Scott, R.J. *et al.* Hereditary nonpolyposis colorectal cancer in 95 families: differences and similarities between mutation-positive and mutation-negative kindreds. *Am. J. Hum. Genet.* **68**, 118–127 (2001).
7. Grindedal, E.M. *et al.* Germ-line mutations in mismatch repair genes associated with prostate cancer. *Cancer Epidemiol. Biomarkers Prev.* **18**, 2460–2467 (2009).
8. Win, A.K. *et al.* Colorectal and other cancer risks for carriers and noncarriers from families with a DNA mismatch repair gene mutation: a prospective cohort study. *J. Clin. Oncol.* **30**, 958–964 (2012).
9. Järvinen, H.J. *et al.* Ten years after mutation testing for Lynch syndrome: cancer incidence and outcome in mutation-positive and mutation-negative family members. *J. Clin. Oncol.* **27**, 4793–4797 (2009).
10. Plon, S.E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).
11. Tavtigian, S.V., Greenblatt, M.S., Goldgar, D.E. & Boffetta, P. Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group. *Hum. Mutat.* **29**, 1261–1264 (2008).
12. Richards, C.S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008).
13. Easton, D.F. *et al.* 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**, 873–883 (2007).
14. Goldgar, D.E. *et al.* Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum. Mutat.* **29**, 1265–1272 (2008).
15. Goldgar, D.E. *et al.* Integrated evaluation of DNA sequence variants of unknown clinical significance: application to *BRCA1* and *BRCA2*. *Am. J. Hum. Genet.* **75**, 535–544 (2004).
16. Thompson, B.A. *et al.* 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. *Hum. Mutat.* **34**, 200–209 (2013).
17. Spurdle, A.B., Couch, F.J., Hogervorst, F.B., Radice, P. & Sinilnikova, O.M. Prediction and assessment of splicing alterations: implications for clinical testing. *Hum. Mutat.* **29**, 1304–1313 (2008).
18. Greenblatt, M.S. *et al.* Locus-specific databases and recommendations to strengthen their contribution to the classification of variants in cancer susceptibility genes. *Hum. Mutat.* **29**, 1273–1281 (2008).
19. Plazzer, J.P. *et al.* The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam. Cancer* **12**, 175–180 (2013).
20. Peltomäki, P. & Vasen, H. Mutations associated with HNPCC predisposition—update of ICG-HNPCC/INSiGHT mutation database. *Dis. Markers* **20**, 269–276 (2004).
21. Peltomäki, P. & Vasen, H.F. 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 (1997).
22. Ou, J. *et al.* Functional analysis helps to clarify the clinical importance of unclassified variants in DNA mismatch repair genes. *Hum. Mutat.* **28**, 1047–1054 (2007).
23. Woods, M.O. *et al.* A new variant database for mismatch repair genes associated with Lynch syndrome. *Hum. Mutat.* **28**, 669–673 (2007).
24. Giardine, B. *et al.* Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nat. Genet.* **43**, 295–301 (2011).
25. Fox, B.I. *et al.* Developing an expert panel process to refine health outcome definitions in observational data. *J. Biomed. Inform.* **46**, 795–804 (2013).
26. Kohonen-Corish, M.R. *et al.* Deciphering the colon cancer genes—report of the InSiGHT–Human Variome Project Workshop, UNESCO, Paris 2010. *Hum. Mutat.* **32**, 491–494 (2011).
27. Thompson, D., Easton, D.F. & Goldgar, D.E. A full-likelihood method for the evaluation of causality of sequence variants from family data. *Am. J. Hum. Genet.* **73**, 652–655 (2003).
28. Senter, L. *et al.* The clinical phenotype of Lynch syndrome due to germ-line *PMS2* mutations. *Gastroenterology* **135**, 419–428 (2008).
29. Baglietto, L. *et al.* Risks of Lynch syndrome cancers for *MSH6* mutation carriers. *J. Natl. Cancer Inst.* **102**, 193–201 (2010).
30. Bonadona, V. *et al.* Cancer risks associated with germline mutations in *MLH1*, *MSH2*, and *MSH6* genes in Lynch syndrome. *J. Am. Med. Assoc.* **305**, 2304–2310 (2011).
31. Mangold, E. *et al.* Spectrum and frequencies of mutations in *MSH2* and *MLH1* identified in 1,721 German families suspected of hereditary nonpolyposis colorectal cancer. *Int. J. Cancer* **116**, 692–702 (2005).
32. Barnetson, R.A. *et al.* Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N. Engl. J. Med.* **354**, 2751–2763 (2006).
33. Patrinos, G.P. *et al.* Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum. Mutat.* **33**, 1503–1512 (2012).
34. Thompson, B.A. *et al.* Calibration of multiple *in silico* tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Hum. Mutat.* **34**, 255–265 (2013).
35. Vallée, M.P. *et al.* Classification of missense substitutions in the *BRCA* genes: a database dedicated to Ex-UVs. *Hum. Mutat.* **33**, 22–28 (2012).
36. Drost, M. *et al.* A rapid and cell-free assay to test the activity of Lynch syndrome–associated *MSH2* and *MSH6* missense variants. *Hum. Mutat.* **33**, 488–494 (2012).
37. Heinen, C.D. & Juel Rasmussen, L. Determining the functional significance of mismatch repair gene missense variants using biochemical and cellular assays. *Hered. Cancer Clin. Pract.* **10**, 9 (2012).
38. Couch, F.J. *et al.* Assessment of functional effects of unclassified genetic variants. *Hum. Mutat.* **29**, 1314–1326 (2008).
39. Rasmussen, L.J. *et al.* Pathological assessment of mismatch repair gene variants in Lynch syndrome: past, present and future. *Hum. Mutat.* **33**, 1617–1625 (2012).
40. Leenen, C.H. *et al.* Pitfalls in molecular analysis for mismatch repair deficiency in a family with biallelic *PMS2* germline mutations. *Clin. Genet.* **80**, 558–565 (2011).
41. Mead, L.J. *et al.* Microsatellite instability markers for identifying early-onset colorectal cancers caused by germ-line mutations in DNA mismatch repair genes. *Clin. Cancer Res.* **13**, 2865–2869 (2007).
42. Plaschke, J. *et al.* Lower incidence of colorectal cancer and later age of disease onset in 27 families with pathogenic *MSH6* germline mutations compared with families with *MLH1* or *MSH2* mutations: the German Hereditary Nonpolyposis Colorectal Cancer Consortium. *J. Clin. Oncol.* **22**, 4486–4494 (2004).
43. Wu, Y. *et al.* Association of hereditary nonpolyposis colorectal cancer–related tumors displaying low microsatellite instability with *MSH6* germline mutations. *Am. J. Hum. Genet.* **65**, 1291–1298 (1999).
44. You, J.F. *et al.* Tumours with loss of MSH6 expression are MSI-H when screened with a pentaplex of five mononucleotide repeats. *Br. J. Cancer* **103**, 1840–1845 (2010).
45. Spurdle, A.B. *et al.* BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk. *J. Med. Genet.* **49**, 525–532 (2012).
46. Xie, J. *et al.* An *MLH1* mutation links *BACH1*/*FANCJ* to colon cancer, signaling, and insight toward directed therapy. *Cancer Prev. Res. (Phila.)* **3**, 1409–1416 (2010).
47. Kosinski, J., Hinrichsen, I., Bujnicki, J.M., Friedhoff, P. & Plotz, G. Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair. *Hum. Mutat.* **31**, 975–982 (2010).
48. Takahashi, M. *et al.* Functional analysis of human MLH1 variants using yeast and *in vitro* mismatch repair assays. *Cancer Res.* **67**, 4595–4604 (2007).
49. Hinrichsen, I. *et al.* Expression defect size among unclassified *MLH1* variants determines pathogenicity in Lynch syndrome diagnosis. *Clin. Cancer Res.* **19**, 2432–2441 (2013).

[1]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [2]School of Medicine, University of Queensland, Brisbane, Queensland, Australia. [3]Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [4]Vermont Cancer Center, University of Vermont College of Medicine, Burlington, Vermont, USA. [5]Division of Molecular Diagnosis and Cancer Prevention, Saitama Cancer Center, Saitama, Japan. [6]Department of Pathology, Faculty of Medicine, Health Sciences Center, Kuwait University, Safat, Kuwait. [7]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [8]Danish Hereditary Non-polyposis Colorectal Cancer Registry, Copenhagen, Denmark. [9]Department of Surgical Gastroenterology, Aalborg University Hospital, Aalborg, Denmark. [10]Hereditary Cancer Program, Catalan Institute of Oncology, Bellvitge Institute for Biomedical Research (IDIBELL), Barcelona, Spain. [11]Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands. [12]Molecular Genetics Laboratory, Victorian Clinical Genetics Services, Murdoch Childrens Research Institute, Melbourne, Victoria, Australia. [13]INSERM UMR S910, Department of Medical Genetics and Functional Genomics, Marseille, France. [14]Department of Cancer Genetics, Mater Private Hospital, Dublin, Ireland. [15]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [16]Institute of Medical Genetics, University Hospital of Wales, Cardiff, UK. [17]INSERM U1079, University of Rouen, Institute for Research and Innovation in Biomedicine, Rouen, France. [18]Department of Genetics, Rouen University Hospital, Rouen, France. [19]Department of Dermatology, University of Utah Medical School, Salt Lake City, Utah, USA. [20]Huntsman Cancer Institute, Salt Lake City, Utah, USA. [21]Center for Molecular Medicine, University of Connecticut Health Center, Farmington, Connecticut, USA. [22]Neag Comprehensive Cancer Center, University of Connecticut Health Center, Farmington, Connecticut, USA. [23]Medizinisch Genetisches Zentrum (MGZ), Munich, Germany. [24]Klinikum der Universität München, Campus Innenstadt, Medizinische Klinik und Poliklinik IV, Munich, Germany. [25]Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [26]School of Medicine, University of Western Sydney, Sydney, New South Wales, Australia. [27]St. Vincent's Clinical School, University of New South Wales, Sydney, New South Wales, Australia. [28]Department of Molecular Medicine and Surgery, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden. [29]Hereditary Gastrointestinal Cancer Genetic Diagnosis Laboratory, Department of Pathology, University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong. [30]Research Group on Inherited Cancer, Department of Medical Genetics, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway. [31]Department of Biosciences, Division of Genetics, University of Helsinki, Helsinki, Finland. [32]Department of Medical Genetics, Haartman Institute, University of Helsinki, Helsinki, Finland. [33]Center for Genetic and Genomic Medicine, First Affiliated Hospital of Zhejiang University School of Medicine, James Watson Institute of Genomic Sciences, Beijing Genome Institute, Beijing, China. [34]School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, New York, USA. [35]Medical Research Council (MRC) Human Genetics Research Unit, Division of Human Genetics, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. [36]Center for Healthy Aging, University of Copenhagen, Copenhagen, Denmark. [37]Institute of Human Genetics, University of Düsseldorf, Düsseldorf, Germany. [38]Discipline of Medical Genetics, Faculty of Health, University of Newcastle, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [39]Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, Australia. [40]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [41]Department of Surgery, State University of New York Downstate, Brooklyn, New York, USA. [42]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada. [43]Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy. [44]Fiorgen Foundation for Pharmacogenomics, Sesto Fiorentino, Florence, Italy. [45]A full list of collaborators assigned microattributions for this study appears at the end of the paper with affiliations. [46]These authors contributed equally to this work. Correspondence should be addressed to M.G. (maurizio.genuardi@unifi.it).

The InSiGHT collaborators are as follows:

Adela Castillejo[47], Adrienne Sexton[48], Anthony K W Chan[29], Alessandra Viel[49], Amie Blanco[50], Amy French[51], Andreas Laner[22], Anja Wagner[52], Ans van den Ouweland[52], Arjen Mensenkamp[53], Artemio Payá[54], Beate Betz[37], Bert Redeker[55], Betsy Smith[56], Carin Espenschied[57], Carole Cummings[58], Christoph Engel[59], Claudia Fornes[60], Cristian Valenzuela[61], Cristina Alenda[54], Daniel Buchanan[62], Daniela Barana[63], Darina Konstantinova[64], Dianne Cairns[65], Elizabeth Glaser[66], Felipe Silva[67], Fiona Lalloo[68], Francesca Crucianelli[44], Frans Hogervorst[69,70], Graham Casey[71], Ian Tomlinson[72], Ignacio Blanco[10], Isabel López Villar[73], Javier Garcia-Planells[74], Jeanette Bigler[75], Jinru Shia[76], Joaquin Martinez-Lopez[77], Johan J P Gille[78], John Hopper[79], John Potter[80], José Luis Soto[47], Jukka Kantelinen[31], Kate Ellis[81], Kirsty Mann[48], Liliana Varesco[82], Liying Zhang[83], Loic Le Marchand[84], Makia J Marafie[85], Margareta Nordling[86], Maria Grazia Tibiletti[87], Mariano Ariel Kahan[88], Marjolijn Ligtenberg[53], Mark Clendenning[62], Mark Jenkins[79], Marsha Speevak[89], Martin Digweed[90], Matthias Kloor[91], Megan Hitchins[92], Megan Myers[50], Melyssa Aronson[93], Mev Dominguez Valentin[94], Michael Kutsche[95], Michael Parsons[1], Michael Walsh[62], Minttu Kansikas[31], Mohd Nizam Zahary[96], Monica Pedroni[97], Nao Heider[98], Nicola Poplawski[99], Nils Rahner[37], Noralane M Lindor[100], Paola Sala[101], Peng Nan[102], Peter Propping[103], Polly Newcomb[80], Rajiv Sarin[104], Robert Haile[71], Robert Hofstra[52], Robyn Ward[92], Rossella Tricarico[44], Ruben Bacares[76], Sean Young[105], Sergio Chialina[60], Serguei Kovalenko[106], Shanaka R Gunawardena[51], Sira Moreno[107], Siu Lun Ho[29], Siu Tsan Yuen[29], Stephen N Thibodeau[51], Steve Gallinger[108], Terrilea Burnett[84], Therese Teitsch[109], Tsun Leung Chan[29], Tom Smyrk[51], Treena Cranston[110], Vasiliki Psofaki[111], Verena Steinke-Lange[37] & Victor-Manuel Barbera[112]

[47]Department of Molecular Genetics, Elche University General Hospital, Elche, Spain. [48]Familial Cancer Centre, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [49]Oncological Referral Center, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Aviano, Italy. [50]Hereditary Gastrointestinal Cancer Prevention Program, University of California, San Francisco, San Francisco, California, USA. [51]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA. [52]Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands. [53]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. [54]Department of Pathology, Hospital Universitario Alicante, Alicante, Spain. [55]Department of Clinical Genetics, Academic Medical Center, Amsterdam, The Netherlands. [56]Benefis Sletten Cancer Institute, Great Falls, Montana, USA. [57]Division of Clinical Cancer Genetics, City of Hope, Duarte, California, USA. [58]Family Cancer Clinic, St. Mark's Hospital, Harrow, UK. [59]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [60]Histocompatibility and Molecular Biology Laboratory, Italian Hospital Garibaldi, Rosario Santa Fe, Argentina. [61]School of Medicine, New York University, New York, New York, USA. [62]Department of Population Health, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [63]Department of Medical Oncology, Health District 5 West Vicenza, Hospital of Montecchio Maggiore, Montecchio Maggiore (VI), Italy. [64]Molecular Medicine Center, Medical University of Sofia, Sofia, Bulgaria. [65]Merseyside and Cheshire Regional Molecular Genetics Laboratory, Liverpool Women's Hospital, Liverpool, UK. [66]International Society for Gastrointestinal Hereditary Tumours. [67]Laboratory of Genomics and Molecular Biology, A.C. Camargo Cancer Center, São Paulo, Brazil.

[68]Manchester Centre for Genomic Medicine, Central Manchester University Hospitals National Health Service (NHS) Foundation Trust, Manchester, UK. [69]Family Cancer Clinic, Netherlands Cancer Institute, Amsterdam, The Netherlands. [70]Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands. [71]Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA. [72]Molecular and Population Genetics Laboratory, London Research Institute, Cancer Research UK, London, UK. [73]Department of Molecular Biology, 12 de Octubre University Hospital, Madrid, Spain. [74]Institute of Genomic Medicine, University of Valencia, Valencia, Spain. [75]Medical Sciences, Amgen, Inc., Seattle, Washington, USA. [76]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [77]Molecular Biology Laboratory, 12 de Octubre University Hospital, Madrid, Spain. [78]Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands. [79]Centre for Molecular Environmental, Genetic and Analytic (MEGA) Epidemiology, University of Melbourne, Melbourne, Victoria, Australia. [80]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [81]Hunter Family Cancer Service, Waratah, New South Wales, Australia. [82]Center for Hereditary Tumours, National Institute for Cancer Research, Genoa, Italy. [83]Diagnostic Molecular Genetics Laboratory, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [84]University of Hawaii Cancer Center, Honolulu, Hawaii, USA. [85]Cancer Genetics Unit, Kuwait Medical Genetics Centre, Kuwait City, Kuwait. [86]Department of Molecular and Clinical Genetics, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [87]Unit of Pathology, Varese Hospital, Varese, Italy. [88]Molecular Oncology Institute of Basic Sciences and Experimental Medicine (ICBME), Hospital Italiano de Buenos Aires, Buenose Aires, Argentina. [89]Division of Genetics, Trillium Health Partners, Credit Valley Hospital, Mississauga, Ontario, Canada. [90]Department of Human Genetics, Institute of Medical Genetics and Human Genetics, Charité–Universitätsmedizin Berlin, Berlin, Germany. [91]Department of Applied Tumor Biology, Institute of Pathology, University of Heidelberg, Heidelberg, Germany. [92]Lowy Cancer Research Centre, Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales, Sydeny, New South Wales, Australia. [93]Familial Gastrointestinal Cancers Unit, Mount Sinai Hospital, Toronto, Ontario, Canada. [94]Department of Oncology, Clinical Science, Lund University, Lund, Sweden. [95]Laboratory for Molecular Medicine, Praenatalzentrum Laboratories, Hamburg, Germany. [96]Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, Pulau Penang, Malaysia. [97]Department of Medicine and Medical Specialties, Modena University Hospital, Modena, Italy. [98]RIKEN Genomic Sciences Center, Yokohama, Japan. [99]South Australia Pathology, Women's and Children's Hospital, Adelaide, South Australia, Australia. [100]Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA. [101]Hereditary Cancers of the Digestive Tract Unit, Predictive and Preventive Medicine, National Tumor Institute IRCCS Foundation, Milan, Italy. [102]School of Life Sciences, Fudan University, Shanghai, China. [103]Institute of Human Genetics, University of Bonn, Bonn, Germany. [104]Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Mumbai, India. [105]Cancer Genetics Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada. [106]Genetic Technologies, Ltd., Melbourne, Victoria, Australia. [107]Genetics Service, Hospital Virgen del Camino, Camino, Spain. [108]Zane Cohen Centre for Digestive Diseases, Toronto, Ontario, Canada. [109]Bioinformatics, Dartmouth Medical School, Dartmouth College, Lebanon, New Hampshire, USA. [110]Oxford Medical Genetics Laboratories, Oxford University Hospitals NHS Trust, Churchill Hospital, Oxford, UK. [111]Biochemical Laboratory, University Hospital of Ioannina, Ioannina, Greece. [112]Research Laboratory, University Hospital of Elche, Elche, Spain.

ONLINE METHODS

InSiGHT Variant Interpretation Committee expertise. The InSiGHT VIC (current chair, M.G.) includes 40 multidisciplinary experts from 5 continents (see **Supplementary Table 9** for disciplines covered by VIC members). The committee is responsible to its Governance Committee, which in turn is responsible to the InSiGHT Council. InSiGHT has recently joined HVP and is a founding member of its Gene and Disease Specific Council. The InSiGHT Council specifically considered the need and responsibility associated with classification assignment on its database and took all reasonable steps to both invite the highest possible expertise to contribute to the classification process and to ensure that its processes and legal standing are robust.

InSiGHT database curation. Mutalyzer[50] was used to standardize the nomenclature of all variants present in the database as of December 2012. Variants with multiple submissions that were originally assigned a classification of pathogenic or probably pathogenic as well as no known pathogenicity or probably no pathogenicity were included in the group of discordant variants used to test the classification criteria. All unique variants identified in the database were assigned to one of the following sources: constitutional, somatic, artificial or unknown.

Development of five-tiered InSiGHT classification criteria. The InSiGHT classification criteria were developed using the Delphi method[25]. A five-tiered classification system originally developed for consistent classification of MMR gene variants identified in participants of the Colon Cancer Family Registry[16,34] was selected as a baseline for the InSiGHT classification criteria. This system included the option of classification on the basis of the posterior probabilities arising from multifactorial likelihood analysis[15,16,51,52] as well as multiple combinations of qualitative data not yet calibrated for inclusion in quantitative analyses but that are often reported in the literature or available from clinical sources. These baseline classification criteria were critically reviewed by InSiGHT VIC members attending the InSiGHT San Antonio meeting in April 2011 and by VIC members via e-mail. In response to comment, the rules were amended for clarity, to apply a more stringent interpretation of functional assay data and to consider additional points of evidence. These InSiGHT rules were used for variant classification over a series of 11 meetings (10 teleconferences and 1 face-to-face meeting), with further changes incorporated after each meeting to include additional points of evidence identified to be relevant during the review process as the committee encountered different combinations of useful data from published and unpublished sources. For example, after discussion, co-occurrence of a variant with a pathogenic mutation in the same gene with clinical information regarding a constitutional MMR deficiency phenotype[53] was included as an *in vivo* test of MMR function, and 1000 Genomes Project data[54] were accepted as a test for population frequency. Consistency of the accumulative evidence required for a given class was reviewed by presentation of the rules at a face-to-face meeting of committee members. Supporting documentation was developed to assist in the interpretation of splicing and functional assay results by B.A.T. in consultation with a subset of committee members with specific expertise in this field (**Fig. 1b** and **Supplementary Tables 4** and **5**). Where necessary, rule alterations were applied retrospectively to variants evaluated in previous meetings. The finalized rules (shown in simplified format in **Fig. 1** and detailed in the **Supplementary Note**) were then used to assess all remaining variants lodged in the InSiGHT database.

Classification of MMR gene variants by literature review and data collation. Variants occurring in the 1000 Genomes Project[54] with allele frequency greater than 1% were automatically classified as class 1 variants. Committee members were invited to participate in at least one classification meeting. A core group participated in each meeting, with attendance invited from VIC membership to make up the balance. Before each meeting, participants were assigned, through randomization, a subset of variants to be assessed. Each attendee was provided literature pertaining to the list of variants to be discussed and, where relevant, additional unpublished clinical or research information submitted by committee members to InSiGHT curator J.-P.P. Meeting attendees were requested to thoroughly review and summarize all information pertaining to the subset of variants in a spreadsheet template and to provide a class assignment based on their interpretation of the information accessed. All reviewer summaries, submitted clinical information and results from causality analysis were compiled into a single file to allow the comparison of data and class assignments for each variant and were circulated to the teleconference participants. During committee meetings, variants were discussed one at a time, assessing the following: class assigned by each reviewer; rationale for classification according to the classification guidelines; difficulties in interpreting specific data sources; assessment of possible redundancy of information due to multiple publications including all or some of the same information pertaining to a variant; differences in interpretation of the guidelines as provided and adjustments required to improve their clarity; the consensus view on variant class considering the preceding discussion; and action required to obtain additional information for refining the classification of variants that remained in class 2, 3 or 4 at the close of discussion. Where classifications differed using qualitative and quantitative criteria, these differences were due to differences in the availability of specific data types for the two approaches, and the most extreme classification was assigned for relevant variants. B.A.T. prioritized variants for examination by identifying and classifying any variants for which rules-based classification could be applied, such as variants that were truncating or comprised a large deletion from nomenclature, canonical splice site with no splicing data or frequency of >1% in a control reference group. B.A.T. then collated all information for all remaining unique variants (including those reviewed previously in teleconferences) and determined which variants had sufficient information to allow classification outside of class 3. Summary information for these variants was circulated for independent class assignment by at least three reviewers from the VIC, and classification was finalized at teleconferences or by e-mail.

Validation of qualitative criteria. A subset of truncating variants and large genomic deletions was selected to validate the qualitative classification criteria. Variants were selected on the basis of the availability of data from the first point of evidence in the qualitative class 5 criterion, i.e., *in vitro* functional assay results (for example, protein truncation test or genomic or mRNA confirmation of large deletions); Constitutional MMR Deficiency Syndrome phenotype; or different haplotypes across multiple families. Published and unpublished data for these variants were then used to validate the other points of evidence required for classification as a class 5 (pathogenic) variant.

Preliminary analysis of class 3 (uncertain) variants. *In silico* probabilities of pathogenicity were estimated for all class 3 missense variants, as described elsewhere[34]. Preliminary bioinformatic analysis of class 3 regulatory variants was undertaken using Encyclopedia of DNA Elements (ENCODE) data[55] on the UCSC Genome Browser.

Implementation of the microattribution process. The variant interpretation process uses both published and unpublished data. For published literature, the PubMed ID (PMID) was used to reference the original work. Some unpublished data were recorded in the InSiGHT database at study initiation, and InSiGHT members were also requested by e-mail to contribute information important for variant classification using a standardized submission template. Data submitters were requested to provide a permanent, publicly searchable unique ID, preferably from the ORCID system, to facilitate the adoption of the microattribution approach. Microattribution was assigned for the different types of information corresponding to the points of evidence required for classification—namely, submitters were allocated one credit of microattribution for each type of information received, including (i) a variant (mandatory), (ii) family history or pedigree, (iii) MSI information, (iv) immunohistochemistry data, (v) *in vitro* functional data, (vi) data from RNA splicing assays and (vii) population frequency data. All unpublished data received by the VIC were recorded in microattribution tables for each element type, with each microattribution table listing a unique researcher ID along with submitted information. Microattribution counts for submitters are publically available on the InSiGHT website. Additionally, the data will be made available in nanopublication format.

Statistical analysis. Multifactorial likelihood analysis was performed for variants with appropriate tumor and segregation data available, using previously

doi:10.1038/ng.2854

reported methods[16,34,51] that are described briefly as follows. Bayes factor analysis was conducted by B.A.T. to assess *MLH1*, *MSH2*, *MSH6* and *PMS2* variant causality from segregation data[16,51] for both published and unpublished pedigrees with sufficient relevant information on cancer and variant carrier status. Penetrance estimates from Senter *et al.*[28] were used in the Bayes segregation analysis[27] of *PMS2* variants. Where family relationship status was unknown, a conservative segregation likelihood ratio was derived—i.e., setting affected carriers as first-degree relatives—which is less informative than segregation between second-degree relatives. Colorectal tumor MSI and somatic *BRAF* mutation status were used to assign likelihood ratios according to tumor phenotype[16]. For each variant, the individual likelihood ratios (cosegregation, tumor) were multiplied to calculate the odds for causality. Then, a posterior probability was calculated by combining the prior probability (*in silico* for missense variants[34] or on the basis of sequence location for all other variants[13]) and the odds for causality using Bayes rule where posterior = (prior × odds × (1/(1−prior)))/(prior × odds × (1/(1 − prior)) + 1). STATA 11 was used to calculate the sample size for the truncating variant validation set, H_0: $P = 0.01$, assuming $\alpha = 0.05$ (one-sided) and power = 0.95.

All other analyses were completed using the statistical package R and GraphPad Prism 6. For meta-analysis of population frequency data, the proportions were combined using an inverse variance random-effects model to account for heterogeneity between studies.

50. Wildeman, M., van Ophuizen, E., den Dunnen, J.T. & Taschner, P.E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum. Mutat.* **29**, 6–13 (2008).
51. Arnold, S. *et al*. Classifying *MLH1* and *MSH2* variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum. Mutat.* **30**, 757–770 (2009).
52. Spurdle, A.B. Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. *Curr. Opin. Genet. Dev.* **20**, 315–323 (2010).
53. Wimmer, K. & Etzler, J. Constitutional mismatch repair–deficiency syndrome: have we so far seen only the tip of an iceberg? *Hum. Genet.* **124**, 105–122 (2008).
54. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
55. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).