# Comprehensive RNA and protein functional assessments contribute to the clinical interpretation of MSH2 variants causing in-frame splicing alterations

Laëtitia Meulemans, Stéphanie Baert Desurmont, Marie-Christine Waill, Gaia Castelain, Audrey Killian, Julie Hauchard, Thierry Frebourg, Florence Coulet, Alexandra Martins, Pascaline Gaildrat, et al.

# Comprehensive RNA and protein functional assessments contribute to the clinical interpretation of MSH2 variants causing in-frame splicing alterations

Laëtitia Meulemans[1], Stéphanie Baert-Desurmont[1,2], Marie-Christine Waill[3], Gaia Castelain[1], Audrey Killian[1], Julie Hauchard[1], Thierry Frebourg[1,2†], Florence Coulet[3,4], Alexandra Martins[1], Martine Muleris[3,4], Pascaline Gaildrat[1,*]

[1] Normandie Univ, UNIROUEN, Inserm U1245, Normandy Centre for Genomic and Personalized Medicine, Rouen, France

[2] Department of Genetics, Rouen University Hospital, Rouen, France

[3] Sorbonne Université, Department of Genetics, AP-HP, Hôpital Pitié-Salpétrière, F-75013 Paris, France

[4] Sorbonne Université, Inserm, Centre de Recherche Saint-Antoine, CRSA, F-75012 Paris, France

[†] Deceased

[*]To whom correspondence should be addressed. Email: pascaline.gaildrat@univ-rouen.fr

**Conflict of interest**

The authors declare that they have no conflict of interest.

**Keywords:** In-frame splicing alterations, *MSH2* gene, Lynch syndrome, *in silico* predictions, minigene splicing assay, methylation tolerance assay, variant clinical interpretation

**Abstract**

**Background** Spliceogenic variants in disease-causing genes are often classified as pathogenic since most of them induce frameshift alterations resulting into protein loss-of-function. However, some of them may induce in-frame anomalies causing protein modifications that may preserve function. Here, we addressed this question by using, as a model system, *MSH2,* a DNA mismatch repair gene implicated in Lynch syndrome.

**Methods** Eighteen *MSH2* variants, mostly localized within canonical splice sites, were characterized by using bioinformatics predictions and splicing minigene assays. The functionality of the resulting protein isoforms was assessed in a methylation tolerance-based assay. Clinicopathological characteristics of variant carriers were collected as well.

**Results** Three in-frame biotypes were identified based on variant-induced spliceogenic outcomes: *(i)* exon skipping (ΔE3, ΔE4, ΔE5, ΔE12), *(ii)* segmental exonic deletion (ΔE7q48, ΔE15p36) and *(iii)* segmental intronic retention (▼E4p24, ▼E7p9, ▼E12q30, ▼E14p9). These splicing events were due to splice site alterations, consistent with computational predictions. The resulting 10 protein isoforms exhibited either *(i)* large deletions (49-93 aa), *(ii)* small deletions (12, 16 aa) or *(iii)* small insertions (3-10 aa) within different functional domains. All these modifications abrogated MSH2 function, in agreement with clinicopathological features of variant carriers.

**Conclusion** These data provide the first functional characterization of in-frame spliceogenic variants in a key MMR gene. They demonstrate that MSH2 function is intolerant to indels triggered by the 18 spliceogenic variants analyzed, supporting their pathogenic nature. This study stresses the importance of combining complementary approaches at both RNA and protein levels to ensure accurate clinical interpretation of in-frame spliceogenic variants.

## Introduction

RNA mis-splicing triggered by specific germline variants is a major cause of human hereditary monogenic disorders [1, 2]. Such spliceogenic impact is generally due to modifications of cis-acting splicing signals, notably 3' and 5' splice sites (3'/5'ss) and auxiliary splicing regulatory elements (SRE) [3]. The consequence of these splicing code alterations is the production of aberrant transcripts either lacking exonic portions or retaining intronic sequences. Most of these splicing anomalies lead to a frameshift of the reference coding sequence, which subsequently results into the introduction of a premature termination codon (PTC). This stop codon gain can then target the aberrant PTC-containing transcripts to selective degradation by a cellular surveillance pathway known as nonsense-mediated mRNA decay (NMD) [4]. If the aberrant RNA escapes NMD, it can then be translated into a truncated protein. Both situations generally cause complete loss-of-function (LoF). Based on this assumption, frameshift-inducing spliceogenic variants are usually considered as null alleles and classified as pathogenic when they are detected in Mendelian disorder-causing genes in which LoF is an established disease mechanism [5].

However, certain spliceogenic variants can also trigger in-frame anomalies. In such circumstances, the possible outcomes at the protein level correspond to either an internal insertion or deletion (indel) of one-to-many amino acids (aa). The functional consequences of such modifications often remain elusive and, consequently, the corresponding variants are usually classified as "variants of uncertain significance" (VUS), hampering optimal clinical management of patients and their relatives [5]. In addition, it is also possible that some specific nucleotide changes initially classified as pathogenic based on their presumed LoF impacts would actually preserve protein function, at least partially, because of their in-frame splicing outcomes, as we recently demonstrated for certain variants in *BRCA2,* one of the major breast and ovarian cancer predisposition genes [6]. This could question the presumed LoF consequences of these variants and subsequently their initial pathogenic classification [5, 6].

In this context, the combined assessment of variant-induced impact on both RNA splicing and protein function emerge as a prerequisite for accurately ascertaining the pathogenic or benign nature of such in-frame spliceogenic variants. In the present study, we addressed this question by using, as a model system, *MSH2* (MIM#609309), a DNA mismatch repair (MMR) gene implicated in Lynch syndrome (LS, MIM#120435), one of the most prevalent hereditary cancer predispositions [7, 8]. Germline monoallelic LoF variants in *MSH2* account for 33% of LS cases [9] and confer markedly increased lifetime risk of developing a spectrum of cancers, primarily colorectal and endometrial cancers, with age-related penetrance and variable expressivity [10–12]. Identification of pathogenic LoF mutations is critical for LS diagnosis and optimization of clinical strategies as well as for identification of at-risk pre-symptomatic relatives who can benefit from risk-reducing management [13].

In this study, we first characterized *MSH2* variants, mostly located at the invariant intronic dinucleotides within the 3'/5'ss (AG/GT, hereafter termed IVS±1/2), leading to in-frame splicing anomalies. The nature and the relative level of variant-induced RNA alterations were assessed by using minigene splicing assays and these experimental data were compared with *in silico* predictions. Then, by taking advantage of a recently developed MSH2 functional assay based on methylation tolerance assessment [14], we evaluated the functionality of MSH2 protein isoforms resulting from these in-frame variant-induced splicing alterations. Altogether, our results contribute to the comprehensive clinical interpretation of in-frame *MSH2* spliceogenic variants.

## Materials and Methods

### Variant selection

*MSH2 v*ariants were described according to the Human Genome Variation Society nomenclature, using the NM_000251.2 reference transcript. *MSH2* variants located within the IVS±1/2 positions flanking in-frame exons were retrieved from national and international human variation databases (Supplemental Materials and Methods) and selected for analysis in splicing minigene assays. In addition, specific in-frame spliceogenic *MSH2* variants, located outside IVS±1/2 of out-of-frame exons, were included in this study based on preliminary data obtained by our group during the course of a systematic splicing impact assessment of *MSH2* variants detected in patients undergoing genetic testing within the French network of oncogenetic laboratories for colorectal cancers (French Genetics and Cancer Group, GGC, Unicancer). These variants as well as natural IVS±1/2 variants that were predicted to have similar splicing impact were selected for further analyses in splicing minigene assays.

### Bioinformatics predictions

Predictions of the variant impact on 3'/5'ss were obtained by using the following *in silico* algorithms: MaxEntScan (MES), SpliceSiteFinder-Like (SSF-L), SPiCE, SpliceAI and SPiP (Supplementary Material and Methods).

### Cell-based minigene splicing assay

In order to assess the variant-induced impact on splicing, we performed splicing assays by using the two-exon minigene pCAS2 vector, as previously described [6, 15, 16], with minor modifications (Supplementary Material and Methods, Supplementary Table S1). This assay is based on the comparative analysis of the splicing patterns of wild-type and mutant minigenes transiently expressed in Hela cells.

**Annotation of splicing events and associated terminology**

Normal exon inclusion is indicated by the symbol +E, followed by the number of the exon. Based on previously published guidelines [17, 18], aberrant transcripts were annotated by using the following symbols: Δ, exonic deletion; ▼, intronic retention; p, 3'ss shift; q, 5'ss shift, followed by the number of nucleotides deleted or inserted. In this manuscript, the term "segmental" is used to describe the deletion of a part of an exon or the retention of a part of an intron whereas the terms "total" or "partial" refer to the extent of the variant-induced splicing impact.

**Methylation tolerance-based functional assay**

In order to assess the functional consequences of the internal aa deletions or insertions resulting from *MSH2* variant-induced in-frame splicing anomalies, we took advantage on the recently developed methylation tolerance-based functional assay [14], with minor modifications (Supplementary Material and Methods). This assay is based on the ability of MMR system to trigger apoptotic signaling in response to DNA damage induced by methylating agents.

**Patient clinical, tumoral and family data**

Clinical, tumoral and family data of patients carrying germline heterozygous *MSH2* variants of interest were collected from the French Universal Mutation Database (UMD-*MSH2*, http://www.umd.be/MSH2/) and in collaboration with the French GGC network. Informed consent for genetic testing was obtained from all participating patients. In addition, information regarding the interpretation of the selected variants, such as patient family history, co-segregation with disease, and tumor characteristics were retrieved from the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) database (https://www.insight-group.org/variants/databases/, last accessed: 05/27/2021).

## Results

### Variant selection

We selected from human variation databases 10 *MSH2* variants based on their location at IVS±1/2 within the 3'/5'ss of in-frame exons (exon 3: c.645+1G>A; exon 4: c.646-2A>G, c.646-1G>C, c.792+1G>A, c.792+1G>C; exon 5: c.942+1_942+2del; exon 12: c.1760-2A>G, c.2005+1G>A, c.2005+2T>C, c.2005+2del). In addition, we included in this study 4 specific *MSH2* variants outside IVS±1/2, located within the 3'/5'ss of out-of-frame exons (exon 7: c.1077-11_1077-7del, c.1276G>A; exon 14: c.2211-10T>G; exon 15: c.2459-3T>G) because these variants were previously identified by our group as in-frame spliceogenic variants during the course of a systematic assessment of the splicing impact of *MSH2* variants detected in patients within the French GGC network. Four natural IVS±1/2 variants at the border of two of these specific out-of-frame exons (exon 7: c.1276+1G>A, c.1276+2T>A; exon 15: c.2459-2A>G, c.2459-1G>T) were also included in this study. Altogether, this selection encompassed a total of 18 *MSH2* variants, including 14 IVS±1/2, 3 intronic variants outside IVS±1/2 and one missense substitution (Supplementary Table S2).

### Characterization of *MSH2* variants responsible for in-frame splicing alterations

In order to experimentally assess the impact on splicing of the 18 selected variants, we took advantage of the pCAS2 minigene splicing assay. On the basis of the in-frame spliceogenic outcomes in this system, we clustered the variants into 3 distinct biotypes: *(i)* exon skipping (7 variants), *(ii)* segmental exonic deletion (6 variants) and *(iii)* segmental intronic retention (5 variants).

#### *MSH2 variants causing in-frame exon skipping*

Within the first biotype, as expected and in agreement with the predicted destruction of the physiological 5' or 3'ss, 7 IVS±1/2 variants induced in-frame exon skipping, albeit to different

extents (Figure 1A, Supplementary Table S3). More precisely, 3 of them, c.645+1G>A (Figure 1B), c.942+1_942+2del (Figure 1C), c.1760-2A>G (Figure 1D), caused total in-frame skipping of exon 3, 5 or 12, respectively (ΔE3, ΔE5, ΔE12 = 100%), whereas the 4 remaining variants, c.792+1G>A, c.792+1G>C (Figure 1E) and c.2005+1G>A, c.2005+2T>C (Figure 1D) were responsible for partial in-frame exon skipping (ΔE4 = 93%, 83% and ΔE12 = 53%, 43%, respectively). Indeed, in the context of the latter variants, extra out-of-frame aberrant transcripts (▼E4q38, ▼E12q31, ▼E12q92) were detected in addition to in-frame exon skipping due to the partial activation of cryptic intronic 5'ss (Figures 1D, 1E) (Supplementary Tables S3 and S4).

***MSH2 variants causing in-frame segmental exonic deletion***

Six of the selected variants belong to the second biotype as they triggered in-frame segmental exonic deletion (Figure 2A). Indeed, one of them is a missense substitution located at the last base of exon 7 (c.1276G>A, p.(Gly426Arg)) that is responsible for a total splicing anomaly resulting into the in-frame deletion of the last 48 nucleotides of exon 7 (ΔE7q48 = 100%) (Figure 2B). This splicing defect is the consequence of the variant-induced weakening of the physiological 5'ss which allows the use of an upstream cryptic exonic 5'ss, in agreement with *in silico* predictions (Figure 2B, Supplementary Table S3). In parallel, we tested two natural IVS±1/2 variants (c.1276+1G>A, c.1276+2T>A) that were predicted to alter the same 5'ss and demonstrated that they both caused identical splicing modifications (Figure 2B, Supplementary Table S3).

Another variant within this category is c.2459-3T>G which caused partial in-frame deletion of the first 36 nucleotides of exon 15 (ΔE15p36 = 52%), as well as two additional aberrant splicing events resulting into frameshift (ΔE15 = 43%, ▼E15p2 = 5%) (Figure 2C, Supplementary Table S4). These 3 aberrant transcripts are the consequence of the variant-induced weakening of the physiological 3'ss, the use of a downstream exonic cryptic 3'ss and the creation of an upstream intronic 3'ss, as indicated by the bioinformatics predictions (Figure 2C, Supplementary Table S3). In parallel, we tested two IVS±1/2 variants that are predicted to alter the same 3'ss (c.2459-2A>G,

c.2459-1G>T) and showed that they induced different relative levels of the in-frame splicing event (ΔE15p36 = 17% and 71%, respectively), as well as an out-of-frame exon skipping (ΔE15 = 83% and 29%, respectively), despite similar *in silico* splice site-dedicated predictions (Figure 2C, Supplementary Tables S3 and S4), suggesting an extra variant effect on SRE.

***MSH2 variants causing in-frame segmental intronic retention***

The splicing impact of the 5 remaining variants referred to a third biotype corresponding to variant-induced in-frame segmental intronic retentions (Figure 3A). Two variants in this category (c.646-2A>G, c.646-1G>C) triggered the retention of the last 27 and 24 nucleotides of intron 3 (▼E4p27 = 55%, 53%; ▼E4p24 = 45%, 47%, respectively) (Figure 3B). These effects are the consequence of the variant-induced destruction of the physiological 3'ss concomitant to the use of two intronic cryptic 3'ss, as indicated by bioinformatics predictions (Figure 3B, Supplementary Table S3). Of note, although the two splicing alterations are in-frame, the ▼E4p27 event leads to the immediate introduction of a PTC (Supplementary Table S4).

Another variant within this category is c.1077-11_1077-7del which caused the major in-frame retention of the last 9 nucleotides of intron 6 (▼E7p9 = 90%), associated with a minor normal inclusion of exon 7 (+E7 = 10%) (Figure 3C). This splicing impact resulted from the drastic variant-induced reduction in strength of the physiological 3'ss and activation of an upstream cryptic intronic 3'ss, as predicted by *in silico* tools (Figure 3C, Supplementary Table S3).

Within this biotype, c.2005+2del led to in-frame retention of the first 30 nucleotides of intron 12 (▼E12q30 = 71%) as well as in-frame skipping of exon 12 (ΔE12 = 25%) and a very minor out-of-frame deletion of the last 92 nucleotides of exon 12 (ΔE12q92 = 4%) (Figure 3D, Supplementary Table S3). These splicing alterations are the consequence of the variant-induced abolition of the physiological 5'ss and the activation of a downstream cryptic intronic 5'ss, predicted by all the algorithms tested, as well as the use of an upstream exonic cryptic 5'ss, only predicted by SSF-L (Figure 3D, Supplementary Table S3).

The last variant belonging to this category is c.2211-10T>G which mainly induced the in-frame insertion of the 9 last nucleotides of intron 13 (▼E14p9 = 67%) and 3 additional minor out-of-frame splicing alterations corresponding to skipping of exon 14 (ΔE14 = 18%) or the deletion of the first 169 or 200 nucleotides of this exon (ΔE14p169 = 6%, ΔE14p200 = 9%) (Figure 3E, Supplementary Tables S3 and S4). The predominant production of ▼E14p9 is due to the variant-induced destruction of the natural 3'ss concomitant to the creation of an upstream intronic 3'ss, whereas ΔE4p169 and ΔE4p200 result from the activation of two downstream cryptic exonic 3'ss, as suggested by the bioinformatics predictions (Figure 3E, Supplementary Table S3).

**Functional analysis of MSH2 protein isoforms resulting from variant-induced in-frame splicing modifications**

Altogether, in the first part of this study, we characterized 18 spliceogenic variants responsible for 3 distinct biotypes of in-frame modifications. These variant-induced splicing alterations would lead to the production of 10 different MSH2 protein isoforms with either *(i)* large internal deletions (49 to 93 aa) (ΔE3: p.Ala123_Gln215del, ΔE4: p.Ile216_Gln264del, ΔE5: p.Val265_Gln314del, ΔE12: p.Tyr588_Gly669del), *(ii)* small internal deletions (12 and 16 aa) (ΔE15p36: p.Gly820_Ala831del, ΔE7q48: p.Ile411_Gly426del) or *(iii)* small internal insertions (3 to 10 aa) (▼E7p9: p.Arg358_Ser359ins3, ▼E14p9: p.Arg737_Ser738ins3, ▼E4p24: p.Glu215_Ile216ins8, ▼E12q30: p.Gly669_Pro670ins10) within different functional domains (Figure 4 and Supplementary Table S4). In order to gain insight into the functionality of these different protein isoforms, we took advantage of a recently published methylation tolerance-based functional assay [14]. Briefly, this test relies on the ability of the MMR system to trigger apoptotic signaling in response to DNA damage induced by methylating agents. Deficient MMR cells do not recognize such lesions and are therefore able to escape apoptosis; they exhibit the so-called methylation tolerance phenotype. Here, we evaluated the ability of the in-frame indel protein isoforms expressed

in LoVo cells, a *MSH2*-deficient cell line, to complement the methylation tolerance phenotype. By using this assay, we demonstrated that all the in-frame deletions or insertions tested, large or small, were unable to restore MMR-dependent DNA damage signaling (Figure 5). Still, one could note that one protein isoform (▼E14p9, p.Arg737_Ser738ins3) exhibited a slightly lower mean survival score, as compared to known pathogenic missense variants (Figure 5).

**Patient clinical, tumoral and family data collected for the selected *MSH2* in-frame spliceogenic variants**

Clinical phenotype, tumoral data and family history were collected within the French GGC network for 40 patients carrying 16 different *MSH2* variants among the 18 in-frame spliceogenic variants selected in this study. All the information has been summarized in Supplementary Table S5. The most frequent cancers reported in patients were colorectal (CRC) and endometrium (EC) cancers with a median age at onset of 49-year-old (range 26-77) and 50 year-old (range 40-75), respectively. All tested tumors from variant carriers except 1 exhibited microsatellite instability (MSI) and/or loss of MSH2 protein expression detected by immunohistochemistry (IHC). In addition, all the information available in the InSiGHT database for 9 reported variants within our selection have been summarized in Supplementary Table S6. These specific variants were classified within this database as pathogenic (Class 5, n = 4) or likely pathogenic (Class 4, n = 5) based on multifactorial likelihood analysis and/or splicing data.

## Discussion

In this study, we provide the first functional characterization of in-frame spliceogenic IVS±1/2 variants in a key MMR gene. By demonstrating that none of the resulting indel protein isoforms retain function, we ascertained the pathogenic nature of these specific variants. Generally, RNA analyzes are performed to demonstrate variant-induced splicing modifications resulting into frameshift [18, 20-22], whereas functional protein assessments are mainly conducted for evaluating the effect of missense variants [14, 23–25]. So far, little attention has been paid to MMR variants responsible for in-frame modifications through mis-splicing, while their functional and clinical consequences may potentially range from disease-causing LoF to neutral impact. Here, we especially focused on *MSH2* IVS±1/2 variants susceptible to have such effects. Within the current consensual framework for variant classification [5, 26], these variants are *a priori* linked to the strongest type of evidence in favor of pathogenicity (PVS1), as they are assumed to trigger LoF. However, caution in the systematic application of this criteria is recommended since some IVS±1/2 variants can lead to in-frame splicing alterations that potentially allow the preservation of protein function. This has been well illustrated by recent works on specific variants in one of the major predisposing breast and ovarian cancer gene, *BRCA2*. Indeed, we and others demonstrated that certain *BRCA2* IVS±1/2 variants, as well as nonsense mutations, trigger in-frame splicing modifications resulting into protein isoforms with retained function [6, 27, 28]. The possible existence of such rescue mechanism might question the pathogenic classification of these specific variants.

Here, we addressed this question by using as a model system another major hereditary cancer gene, *MSH2*. In this study, by assessing the splicing impact of 18 variants including 14 IVS±1/2 alterations, we characterized 3 distinct biotypes of in-frame spliceogenic *MSH2* variants, on the basis of their RNA and protein outcomes corresponding to either *(i)* exon skipping resulting into large protein internal deletions (del 49 to 93 aa), *(ii)* segmental exonic deletions resulting into small

protein internal deletions (del 12 and 16 aa) or *(iii)* segmental intronic retentions resulting into protein small internal insertions (ins 3 to 10 aa) (Supplementary Figure S1).

All these in-frame mis-splicing events are the consequence of variant-induced disruption of the physiological 5'/3'ss, alone or combined with the activation of cryptic exonic or intronic 5'/3'ss, in agreement with bioinformatics predictions. Yet, only the experimental minigene-based approach enabled the assessment of the relative proportions of the different RNA isoforms produced in the mutant context. Among the 18 selected variants, 12 induced the production of multiple aberrant in-frame or frameshift transcripts, with up to 4 distinct RNA isoforms detected in the context of one specific variant. These observations underline the importance of the exhaustive and quantitative experimental evaluation of mis-splicing for accurately infer the variant biological impact. Such information is not always available from patient RNA analyses. Still, it is to note that seven of the 18 variant-induced in-frame mis-splicing detected by using minigene assays in this study have been previously observed in patient RNA samples (Supplementary Table S7), confirming the biological relevance of the minigene splicing reporter system, as previously shown for MMR genes [21, 22, 29].

We took advantage of the methylation tolerance-based assay to assess the functional consequences of the different *MSH2* in-frame spliceogenic variants. This approach examines one of the two key MMR functions by measuring the cellular ability to activate apoptosis in response to DNA damage [30, 31]. Recent work demonstrated that it exhibits 95% sensitivity and 100% specificity in discriminating between *MSH2/MLH1* missense pathogenic and neutral variants [14], indicating that this well-calibrated system is a pertinent method to gain insight into variant interpretation. By using this assay, we demonstrated that all the 10 MSH2 indel protein isoforms resulting from the different in-frame spliceogenic variants were defective in MMR function. Still, it is to note that one of these protein isoforms (p.Arg737_Ser738ins3 resulting from c.2211-10T>G) displays a mean survival

score that is slightly lower as compared to the known pathogenic variants of the calibration set. Further analyses will be required to confirm that this specific variant is responsible for total LoF.

To our knowledge, only two of these 10 indel protein isoforms (ΔE5: p.Val265_Gln314del, ▼E4p24: p.Glu215_Ile216ins8) have been previously tested in a functional assay based on stable murine/human hybrid cell lines, with concordant results (*i.e.* defective MMR activity) [32]. Indirectly, it was already speculate that ΔE5 protein isoform was dysfunctional as the most frequent of all known pathogenic *MSH2* variants (*i.e.* c.942+3A>T) is indeed a spliceogenic variant leading to the in-frame skipping of this specific exon [33]. In addition, it should be pointed out that ΔE5 as well as ΔE3 have been previously reported as naturally occurring alternative transcripts [18]. Because we demonstrated that these specific alternative in-frame splicing events result into MSH2 LoF, we can now exclude the possibility that *MSH2* ΔE5 and ΔE3 could represent "rescue" isoforms, in contrast to what has been suggested for *BRCA1* ΔE9-10 [34].

Altogether, our data demonstrate that MSH2 function does not tolerate any of the large (≥49 aa) nor even small (≤16 aa) in-frame indels caused by the spliceogenic variants characterized in this study. In contrast, it has been recently shown by using deep mutational scanning combined with massively parallel functional MMR assay that MSH2 exhibits a remarkable degree of tolerance to missense variants [25]. Indeed, despite the high degree of evolutionary sequence conservation, the large majority (89%) of the ~ 17 000 single amino acid substitutions tested within this protein were functionally neutral. The LoF caused by in-frame spliceogenic variants could be the consequence of the resulting protein indels within specific MSH2 essential functional domains [35–38], as each of them might need to be fully intact to be active. Under this hypothesis, defective MMR capacity would be the result of the failure of one of the multiple specific activities supporting MSH2 function, such as protein-protein interaction (MSH2–MSH6 complex formation), DNA binding (mismatch recognition) or enzymatic activity (ATP binding/hydrolysis). Another explanation would be that these in-frame modifications impaired MSH2 function by conferring a defect in protein

folding and structural integrity leading eventually to protein destabilization and degradation. This second hypothesis might be favored based on the observed loss of MSH2 expression detected by IHC in the tumors of most patients bearing the in-frame spliceogenic variants analyzed in our study. In addition, it has been previously shown that 2 of the 10 indel protein isoforms tested here (ΔE5: p.Val265_Gln314del; ▼E4p24: p.Glu215_Ile216ins8) lead to the absence of MSH2 protein detection in murine/human hybrid cell lines [32]. Still, we cannot exclude the possibility that other *MSH2* variants causing in-frame modifications can allow the production of protein isoforms that may conserve activity. Interestingly, it has been demonstrated that the loss of the translation start codon induced by *MSH2* c.1A>C triggers the use of a downstream in-frame alternative AUG leading to the production of a protein deleted for its first 25 aa [39–41]. This N-terminal truncated protein isoform retains function, and its expression may be responsible for the reduced penetrance observed for this variant. Based on these data, it is conceivable that certain deletions within the N-terminal region of *MSH2* may preserve protein functionality.

Altogether, our results should contribute to the accurate clinical classification of the *MSH2* in-frame spliceogenic variants characterized in this study. By ascertaining their LoF status, our data support the initial pathogenic classification of the 14 in-frame spliceogenic IVS±1/2 variants, consistent with patient clinical, tumoral and family data. Moreover, this interpretation can, in principle, be extended to all *MSH2* genomic deletions reported in mutational databases as responsible for the same types of in-frame modifications. Our data also support the pathogenic classification of 4 additional in-frame spliceogenic variants located outside IVS±1/2 (3 intronic and 1 missense) which were reported with conflicting interpretations in databases (Supplementary Table S2).

The strength of our strategy was to assess the variant impact at both RNA and protein levels by using complementary approaches. By design, our study framework required to sequentially evaluate the variant-induced splicing modifications and their protein functional consequences into two separate independent assays. In the future, it would be profitable to gain insight into the

spliceogenic variant in-frame impact on both RNA splicing and protein function in a unique integrated system relying on the CRISPR-Cas9-mediated introduction of the variant into the endogenous *MSH2* chromosomal locus, as recently described for *MSH2* missense variants [24].

In summary, these data provide the first combined characterization on both RNA splicing and protein function of in-frame spliceogenic variants in a key MMR gene. They demonstrate that MSH2 function is intolerant to in-frame indels triggered by the 18 spliceogenic variants analyzed, supporting their pathogenic nature. This study emphasizes the importance of the comprehensive elucidation of the variant-induced biological effect on RNA splicing and the resulting consequence on protein function to ensure accurate clinical interpretation of in-frame spliceogenic variants.

**Acknowledgements**

This paper is dedicated to the memory of Professor Thierry Frebourg.

The authors thank: all clinical and molecular geneticists members of the French oncogenetic network for colorectal cancers (Group "Genetic and Cancer", GGC, Unicancer) for reporting variant information and patient clinical and tumoral data in the UMD database, Drs M.-P. Buisine, C. Maugard, L. Goldmard, C. Popovici for providing additional patient clinical and tumoral data, Drs S. Olschwang, C. Dugast, J.-M. Rey, M. Guillaud-Bataille, P. Laurent-Puig for sending patient genomic DNA. We are also grateful to the Flow Cytometry core (CYTO-ICAN) of ICAN institute for the flow cytometry analysis. This work was financially supported by the Fédération Hospitalo-Universitaire (FHU) Normandy Centre for Genomic and Personalized Medicine (NGP), the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (Gefluc, # R18064EE), a translational research grant from the French National Cancer Institute and the Direction Generale de l'Offre des Soins (INCa/DGOS, AAP/CFB/CI, FASDEC), as well as the European Union and Région Normandie. Europe gets involved in Normandie with the European Regional Development Fund (ERDF). L. Meulemans was funded by a FHU-NGP PhD fellowship and benefitted from two short-term mobility fellowships, one from Cancéropôle Nord-Ouest (CNO) and the other from EDnBISE.

## References


1 Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet* 2016;17:19–32.

2 Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet* 2017;49:848–55.

3 Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002;3:285–98.

4 Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* 2019;20:406–20.

5 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, on behalf of the ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–23.

6 Meulemans L, Mesman RLS, Caputo SM, Krieger S, Guillaud-Bataille M, Caux-Moncoutier V, Léone M, Boutry-Kryza N, Sokolowska J, Révillion F, Delnatte C, Tubeuf H, Soukarieh O, Bonnet-Dorion F, Guibert V, Bronner M, Bourdon V, Lizard S, Vilquin P, Privat M, Drouet A, Grout C, Calléja FMGR, Golmard L, Vrieling H, Stoppa-Lyonnet D, Houdayer C, Frebourg T, Vreeswijk MPG, Martins A, Gaildrat P. Skipping Nonsense to Maintain Function: The Paradigm of *BRCA2* Exon 12. *Cancer Res* 2020;80:1374–86.

7 Lynch HT, de la Chapelle A. Hereditary Colorectal Cancer. *N Engl J Med* 2003;48:919-932.

8 Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. *Nat Rev Cancer* 2015;15:181–94.

9 Plazzer JP, Sijmons RH, Woods MO, Peltomäki P, Thompson B, Den Dunnen JT, Macrae F. The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam Cancer* 2013;12:175–80.

10 Møller P, Seppälä TT, Bernstein I, Holinski-Feder E, Sala P, Gareth Evans D, Lindblom A, Macrae F, Blanco I, Sijmons RH, Jeffries J, Vasen HFA, Burn J, Nakken S, Hovig E, Rødland EA, Tharmaratnam K, de Vos tot Nederveen Cappel WH, Hill J, Wijnen JT, Jenkins MA, Green K, Lalloo F, Sunde L, Mints M, Bertario L, Pineda M, Navarro M, Morak M, Renkonen-Sinisalo L, Valentin MD, Frayling IM, Plazzer J-P, Pylvanainen K, Genuardi M, Mecklin J-P, Moeslein G, Sampson JR, Capella G. Cancer risk and survival in *path_MMR* carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database. *Gut* 2018;67:1306–16.

11 ten Broeke SW, van Bavel TC, Jansen AML, Gómez-García E, Hes FJ, van Hest LP, Letteboer TGW, Olderode-Berends MJW, Ruano D, Spruijt L, Suerink M, Tops CM, van Eijk R, Morreau H, van Wezel T, Nielsen M. Molecular Background of Colorectal Tumors From Patients With Lynch Syndrome Associated With Germline Variants in PMS2. *Gastroenterology* 2018;155:844–51.

12 Dominguez-Valentin M, Sampson JR, Seppälä TT, ten Broeke SW, Plazzer J-P, Nakken S, Engel C, Aretz S, Jenkins MA, Sunde L, Bernstein I, Capella G, Balaguer F, Thomas H, Evans DG, Burn J, Greenblatt M, Hovig E, de Vos tot Nederveen Cappel WH, Sijmons RH, Bertario L, Tibiletti MG, Cavestro GM, Lindblom A, Della Valle A, Lopez-Köstner F, Gluck N, Katz LH, Heinimann K, Vaccaro CA, Büttner R, Görgens H, Holinski-Feder E, Morak M, Holzapfel S, Hüneburg R, Knebel Doeberitz M von, Loeffler M, Rahner N, Schackert HK, Steinke-Lange V, Schmiegel W, Vangala D, Pylvänäinen K, Renkonen-Sinisalo L, Hopper JL, Win AK, Haile RW, Lindor NM, Gallinger S, Le Marchand L, Newcomb PA, Figueiredo JC, Thibodeau SN, Wadt K, Therkildsen C, Okkels H, Ketabi Z, Moreira L, Sánchez A, Serra-Burriel M, Pineda M, Navarro M, Blanco I, Green K, Lalloo F, Crosbie EJ, Hill J, Denton OG, Frayling IM, Rødland EA, Vasen H, Mints M, Neffa F, Esperon P, Alvarez K, Kariv R, Rosner G, Pinero TA, Gonzalez ML, Kalfayan P, Tjandra D, Winship IM, Macrae F, Möslein G, Mecklin J-P, Nielsen M, Møller P. Cancer risks by gene, age, and gender in 6350 carriers ofpathogenic mismatch repair variants: findings from the Prospective Lynch SyndromeDatabase. *Genet Med* 2020;22:15–25.

13 Vasen HFA, Blanco I, Aktan-Collan K, Gopie JP, Alonso A, Aretz S, Bernstein I, Bertario L, Burn J, Capella G, Colas C, Engel C, Frayling IM, Genuardi M, Heinimann K, Hes FJ, Hodgson SV, Karagiannis JA, Lalloo F, Lindblom A, Mecklin J-P, Møller P, Myrhoj T, Nagengast FM, Parc Y, Ponz de Leon M, Renkonen-Sinisalo L, Sampson JR, Stormorken A, Sijmons RH, Tejpar S, Thomas HJW, Rahner N, Wijnen JT, Järvinen HJ, Möslein G. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut* 2013;62:812–23.

14 Bouvet D, Bodo S, Munier A, Guillerm E, Bertrand R, Colas C, Duval A, Coulet F, Muleris M. Methylation Tolerance-Based Functional Assay to Assess Variants of Unknown Significance in the MLH1 and MSH2 Genes and Identify Patients With Lynch Syndrome. *Gastroenterology* 2019;157:421–31.

15 Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, Tosi M. Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. *Methods Mol Biol Clifton NJ* 2010;653:249–57.

16 Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frébourg T, Tosi M, Martins A. Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLoS Genet* 2016;12:e1005756.

17 Colombo M, Blok MJ, Whiley P, Santamariña M, Gutiérrez-Enríquez S, Romero A, Garre P, Becker A, Smith LD, De Vecchi G, Brandão RD, Tserpelis D, Brown M, Blanco A, Bonache S, Menéndez M, Houdayer C, Foglia C, Fackenthal JD, Baralle D, Wappenschmidt B, kConFaB Investigators, Díaz-Rubio E, Caldés T, Walker L, Díez O, Vega A, Spurdle AB, Radice P, De La Hoya M. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Hum Mol Genet* 2014;23:3666–80.

18 Thompson BA, Martins A, Spurdle AB. A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays: A review of mismatch repair gene transcripts. *Clin Genet* 2015;87:100–8.

19 Lützen A, de Wind N, Georgijevic D, Nielsen FC, Rasmussen LJ. Functional analysis of HNPCC-related missense mutations in MSH2. *Mutat Research* 2008;645:44–55

20 Auclair J, Busine MP, Navarro C, Ruano E, Montmain G, Desseigne F, Saurin JC, Lasset C, Bonadona V, Giraud S, Puisieux A, Wang Q. Systematic mRNA analysis for the effect ofMLH1 andMSH2 missense and silent mutations on aberrant splicing. *Hum Mutat* 2006;27:145–54.

21 Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J, Frébourg T, Tosi M. A large fraction of unclassified variants of the mismatch repair genes *MLH1* and *MSH2* is associated with splicing defects. *Hum Mutat* 2008;29:1412–24.

22 Thompson BA, Walters R, Parsons MT, Dumenil T, Drost M, Tiersma Y, Lindor NM, Tavtigian SV, de Wind N, Spurdle AB, the InSiGHT Variant Interpretation Committee. Contribution of mRNA Splicing to Mismatch Repair Gene Sequence Variant Interpretation. *Front Genet* 2020;11:798.

23 Drost M, Tiersma Y, Thompson BA, Frederiksen JH, Keijzers G, Glubb D, Kathe S, Osinga J, Westers H, Pappas L, Boucher KM, Molenkamp S, Zonneveld JB, van Asperen CJ, Goldgar DE, Wallace SS, Sijmons RH, Spurdle AB, Rasmussen LJ, Greenblatt MS, de Wind N, Tavtigian SV. A functional assay–based procedure to classify mismatch repair gene variants in Lynch syndrome. *Genet Med* 2019;21:1486–96.

24 Rath A, Mishra A, Ferreira VD, Hu C, Omerza G, Kelly K, Hesse A, Reddi HV, Grady JP, Heinen CD. Functional interrogation of Lynch syndrome-associated *MSH2* missense variants via CRISPR-Cas9 gene editing in human embryonic stem cells. *Hum Mutat* 2019;40:2044–56.

25 Jia X, Burugula BB, Chen V, Lemons RM, Jayakody S, Maksutova M, Kitzman JO. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet* 2021;108:163–75.

26 Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, Harrison SM, ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* 2018;39:1517–24.

27 Stauffer S, Biswas K, Sharan SK. Bypass of premature stop codons and generation of functional BRCA2 by exon skipping. *J Hum Genet* 2020;65:805–9.

28 Mesman RLS, Calléja FMGR, de la Hoya M, Devilee P, van Asperen CJ, Vrieling H, Vreeswijk MPG. Alternative mRNA splicing can attenuate the pathogenicity ofpresumed loss-of-function variants in BRCA2. *Genet Med* 2020;22:1355–65.

29 van der Klift HM, Jansen AML, van der Steenstraten N, Bik EC, Tops CMJ, Devilee P, Wijnen JT. Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. *Mol Genet Genomic Med* 2015;3:327–45.

30 Heinen CD. Mismatch repair defects and Lynch syndrome: The role of the basic scientist in the battle against cancer. *DNA Repair* 2016;38:127–34.

31 Li Z, Pearlman AH, Hsieh P. DNA mismatch repair and the DNA damage response. *DNA Repair* 2016;38:94–101.

32 Marra G, D'Atri S, Yan H, Perrera C, Cannavo' E, Vogelstein B, Jiricny J. Phenotypic Analysis of hMSH2 Mutations in Mouse Cells Carrying Human Chromosomes. *Cancer Res* 2001;61:7719-21.

33 Desai DC, Lockman JC, Chadwick RB, Gao X, Percesepe A, Evans DG, Miyaki M, Yuen ST, Radice P, Maher ER, Wright FA, de La Chapelle A. Recurrent germline mutation in MSH2 arises frequently de novo. *J Med Genet* 2000;37:646–52.

34 de la Hoya M, Soukarieh O, López-Perolio I, Vega A, Walker LC, van Ierland Y, Baralle D, Santamariña M, Lattimore V, Wijnen J, Whiley P, Blanco A, Raponi M, Hauke J, Wappenschmidt B, Becker A, Hansen TVO, Behar R, Investigators Kc, Niederacher D, Arnold N, Dworniczak B, Steinemann D, Faust U, Rubinstein W, Hulick PJ, Houdayer C, Caputo SM, Castera L, Pesaran T, Chao E, Brewer C, Southey MC, van Asperen CJ, Singer CF, Sullivan J, Poplawski N, Mai P, Peto J, Johnson N, Burwinkel B, Surowy H, Bojesen SE, Flyger H, Lindblom A, Margolin S, Chang-Claude J, Rudolph A, Radice P, Galastri L, Olson JE, Hallberg E, Giles GG, Milne RL, Andrulis IL, Glendon G, Hall P, Czene K, Blows F, Shah M, Wang Q, Dennis J, Michailidou K, McGuffog L, Bolla MK, Antoniou AC, Easton DF, Couch FJ, Tavtigian S, Vreeswijk MP, Parsons M, Meeks HD, Martins A, Goldgar DE, Spurdle AB. Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum Mol Genet* 2016;25:2256–68.

35 Lamers MH, Perrakis A, Enzlin JH, Winterwerp HHK, de Wind N, Sixma TK. The crystal structure of DNA mismatch repair protein MutS binding to a G·T mismatch. *Nature* 2000;407:711–7.

36 Warren JJ, Pohlhaus TJ, Changela A, Iyer RR, Modrich PL, Beese LS. Structure of the Human MutSα DNA Lesion Recognition Complex. *Mol Cell* 2007;26:579–92.

37 Lützen A, de Wind N, Georgijevic D, Nielsen FC, Rasmussen LJ. Functional analysis of HNPCC-related missense mutations in MSH2. *Mutat Res Mol Mech Mutagen* 2008;645:44–55.

38 Edelbrock MA, Kaliyaperumal S, Williams KJ. Structural, molecular and cellular functions of MSH2 and MSH6 during DNA mismatch repair, damage signaling and other noncanonical activities. *Mutat Res Mol Mech Mutagen* 2013;743–744:53–66.

39 Kets CM, Hoogerbrugge N, van Krieken JHJM, Goossens M, Brunner HG, Ligtenberg MJL. Compound heterozygosity for two MSH2 mutations suggests mild consequences of the initiation codon variant c.1A>G of MSH2. *Eur J Hum Genet* 2009;17:159–64.

40 Cyr JL, Brown GD, Stroop J, Heinen CD. The predicted truncation from a cancer-associated variant of the *MSH2* initiation codon alters activity of the MSH2-MSH6 mismatch repair complex. *Mol Carcinog* 2012;51:647–58.

41 Rosenthal ET, Bowles KR, Pruss D, van Kan A, Vail PJ, McElroy H, Wenstrup RJ. Exceptions to the rule: Case studies in the prediction of pathogenicity for genetic variants in hereditary cancer genes. *Clin Genet* 2015;88:533–41.

**Legends of the Figures**

**Figure 1. Spliceogenic *MSH2* variants causing in-frame exon skipping**

**A. Distribution of the 7 selected variants**. The figure represents the structure of the *MSH2* gene with the phasing of its 16 exons indicated by the rounded, square and pointed shapes (see legend embedded in the figure). In-frame exons are further highlighted in black. The variants of interest are located at the canonical intronic dinucleotides (IVS±1/2) within the 3' or 5'ss of these specific exons. **B-E. Variant-induced impact on splicing of exon 3 (B), exon 4 (C), exon 5 (D) and exon 12 (E), as determined in a cell-based minigene splicing assay.** Wild-type (WT) and mutant pCAS2 minigene constructs carrying the exon of interest were transiently expressed in HeLa cells and the splicing patterns of the transcripts produced from the different minigenes were analyzed by semi-quantitative fluorescent RT-PCR, as indicated in Supplementary Materials and Methods. The relative levels of the different RNA isoforms were quantified by resolving the fluorescent RT-PCR products by capillary electrophoresis under denaturing conditions on an automated 3500 Genetic Analyzer (Applied Biosystems). One representative electropherogram analyzed by using the GeneMapper v5.0 Software (Applied Biosystems) is shown for each variant and WT constructs. The fluorescence peaks correspond to the different RT-PCR products resolved according to their respective sizes. The identities of the different transcripts are indicated on the top of each panel: transcripts with normal exon inclusion (+E), transcripts with whole or segmental deletions of an exon (ΔE) and transcripts with intronic retentions (▼E). The symbols p and q represent respectively 3' and 5'ss shift, followed by the number of nucleotides deleted or inserted. The RNA isoforms indicated in black are in-frame whereas those reported in grey are out-of-frame. Peak areas were used to quantify the relative levels of each transcript, expressed as percentages, and correspond to the average of at least three independent experiments. On the left, the splicing patterns of the physiological and aberrant transcripts are schematically represented. On the right, the *in silico* MES predictions of the physiological (Ψ) or cryptic (c) 3'/5'ss are shown in the WT and variant contexts.

**Figure 2. Spliceogenic *MSH2* variants causing in-frame segmental exonic deletion**

**A. Distribution of the 6 selected variants**. The figure shows the schematic representation of the *MSH2* gene with the phasing of its 16 exons, as already described in Figure 1A. The variants of interest are located at the 3' or 5'ss of out-of-frame exons indicated in dark grey. **B-C. Variant-induced impact on splicing of exon 7 (B) and exon 15 (C), as determined in cell-based minigene splicing assays.** Legends as described in Figure 1B-E. When transcripts were close in size, stars (*) were used to link RNA isoforms with their corresponding relative levels. On the left,

the splicing patterns of the physiological and aberrant transcripts are schematically represented. On the right, the *in silico* MES predictions of the physiological (Ψ), new (n) or cryptic (c) 3'/5'ss are shown in the WT and variant contexts.

**Figure 3. Spliceogenic *MSH2* variants causing in-frame segmental intronic retention**

**A. Distribution of the 5 selected variants**. The figure shows the schematic representation of the *MSH2* gene with the phasing of its 16 exons, as explained in Figure 1A. The variants of interest are located at the 3' or 5'ss of in-frame or out-of-frame exons indicated, respectively, in black and dark grey. **B-E. Variant-induced impact on splicing of exon 4 (B), exon 7 (C), exon 12 (D) and exon 14 (E), as determined in cell-based minigene splicing assays.** Legends as described in Figure 1.B-E.

**Figure 4. Schematic representation of the variant-induced in-frame splicing modifications at the RNA level and the corresponding consequences at the MSH2 protein level.**

The top panel shows the schematic representation of the *MSH2* mRNA with its 16 exons. In-frame exons are indicated in black. The different variant-induced in-frame splicing anomalies are indicated, corresponding to either *(i)* exon skipping (ΔE), *(ii)* segmental exonic deletion (ΔEp/q) or *(iii)* segmental intronic retention (▼Ep/q). The symbols p and q represent 3' and 5'ss shift, respectively, followed by the number of nucleotides deleted or inserted. The bottom panel shows the schematic representation of the MSH2 protein with its different functional domains, as reported by [19]. The respective consequences at the protein level of the different in-frame splicing anomalies are indicated, corresponding to either *(i)* large internal deletion (del 49 to 93 aa, 4 protein isoforms), *(ii)* small internal deletion (del 12 and 16 aa, 2 protein isoforms) or *(iii)* small internal insertion (ins 3 to 10 aa, 4 protein isoforms) within different functional domains.

**Figure 5. Functional analysis of MSH2 protein isoforms resulting from variant-induced in-frame splicing alterations**

A total of 10 MSH2 protein isoforms were tested in methylation tolerance-based functional assays, in parallel to known pathogenic (n=10) or neutral (n=10) missense variants, used as controls. The 10 MSH2 protein isoforms resulting from variant-induced in-frame splicing alterations can be divided into 3 distinct biotypes: *(i)* exon skipping resulting into protein large internal deletions (n=4), *(ii)* segmental exonic deletions resulting into protein small internal deletions (n=2) or *(iii)*

segmental intronic retentions resulting into protein small internal insertions (n=4). The mean survival score represents the average survival fraction after 1 and 2 treatments with 1 μM MNNG. The gray area represents the interval between the values of the 2 validation sets (pathogenic set and neutral set). The cutoff value as determined by the ROC curve in the validation set is represented by a dotted line.

**Figure 1**

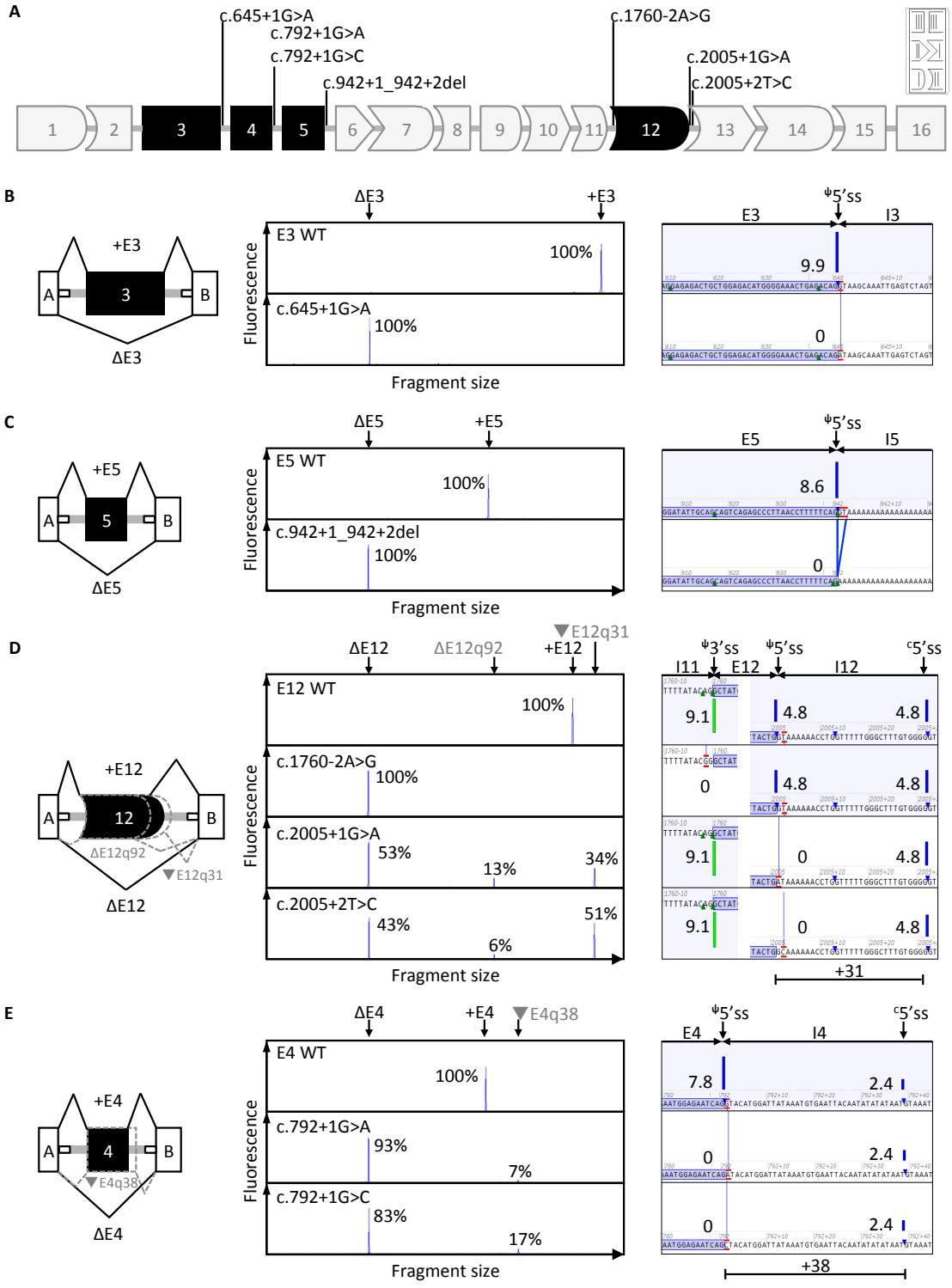

A
c.645+1G>A
c.792+1G>A
c.792+1G>C
c.942+1_942+2del
c.1760-2A>G
c.2005+1G>A
c.2005+2T>C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
B
+E3
A
3
B
ΔE3
ΔE3
+E3
E3 WT
100%
c.645+1G>A
100%
Fluorescence
Fragment size
Ψ5'ss
E3
I3
9.9
0
C
+E5
A
5
B
ΔE5
ΔE5
+E5
E5 WT
100%
c.942+1_942+2del
100%
Fluorescence
Fragment size
Ψ5'ss
E5
I5
8.6
0
D
+E12
A
12
B
ΔE12q92
▼E12q31
ΔE12
ΔE12
ΔE12q92
+E12
▼E12q31
E12 WT
100%
c.1760-2A>G
100%
c.2005+1G>A
53%
13%
34%
c.2005+2T>C
43%
6%
51%
Fluorescence
Fragment size
Ψ3'ss
Ψ5'ss
c5'ss
I11
E12
I12
9.1
4.8
4.8
0
4.8
4.8
9.1
0
4.8
9.1
0
4.8
+31
E
+E4
A
4
B
▼E4q38
ΔE4
ΔE4
+E4
▼E4q38
E4 WT
100%
c.792+1G>A
93%
7%
c.792+1G>C
83%
17%
Fluorescence
Fragment size
Ψ5'ss
c5'ss
E4
I4
7.8
2.4
0
2.4
0
2.4
+38

**Figure 2**

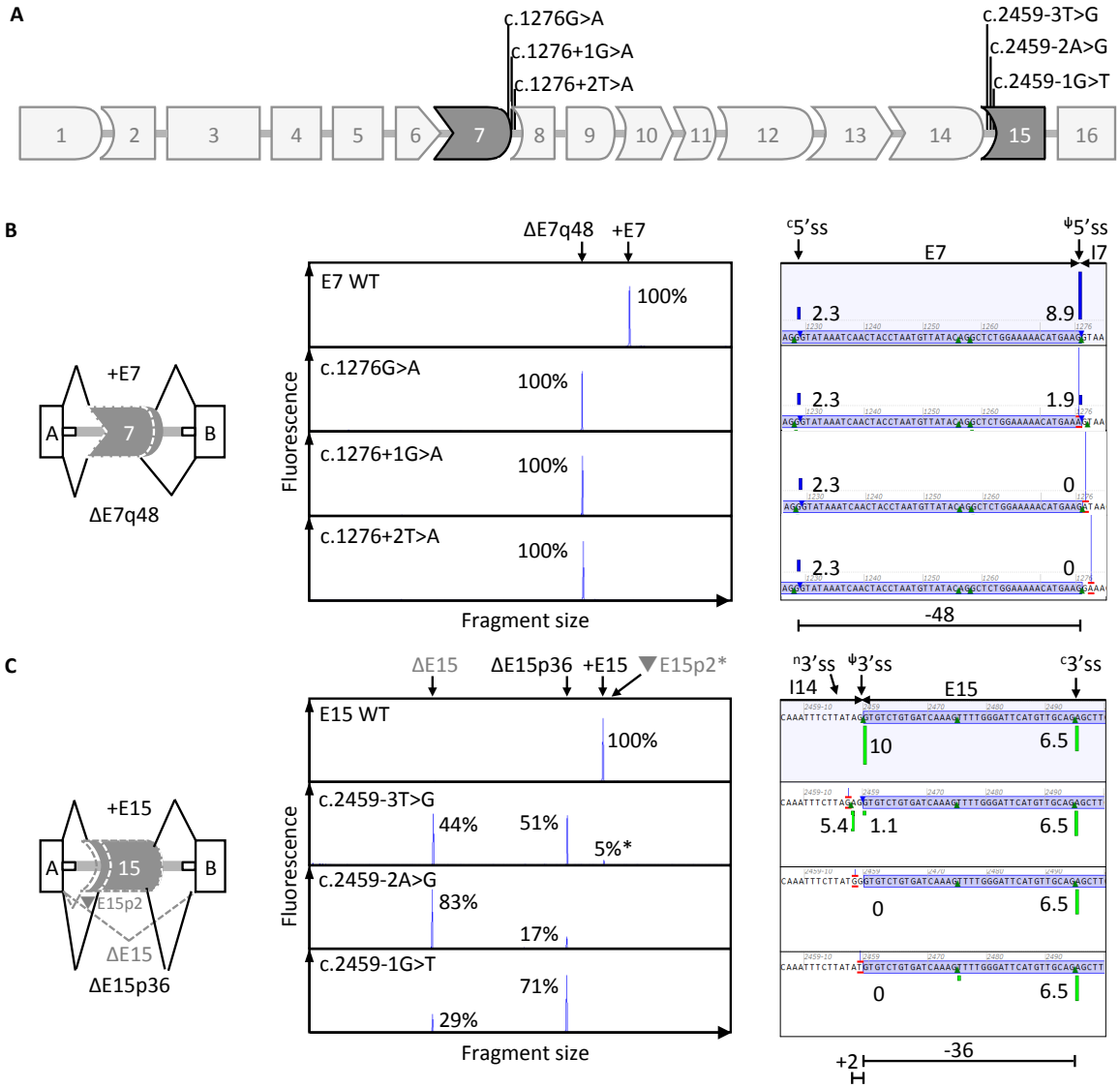

A
c.1276G>A
c.1276+1G>A
c.1276+2T>A
c.2459-3T>G
c.2459-2A>G
c.2459-1G>T
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
B
+E7
A
7
B
ΔE7q48
ΔE7q48 +E7
E7 WT
100%
c.1276G>A
100%
c.1276+1G>A
100%
c.1276+2T>A
100%
Fluorescence
Fragment size
c5'ss
Ψ5'ss
E7
I7
2.3
8.9
2.3
1.9
2.3
0
2.3
0
-48
C
+E15
A
15
B
E15p2
ΔE15
ΔE15p36
ΔE15 ΔE15p36 +E15 E15p2*
E15 WT
100%
c.2459-3T>G
44%
51%
5%*
c.2459-2A>G
83%
17%
c.2459-1G>T
71%
29%
Fluorescence
Fragment size
n3'ss
Ψ3'ss
c3'ss
I14
E15
10
6.5
5.4
1.1
6.5
0
6.5
0
6.5
+2
-36

**Figure 3**

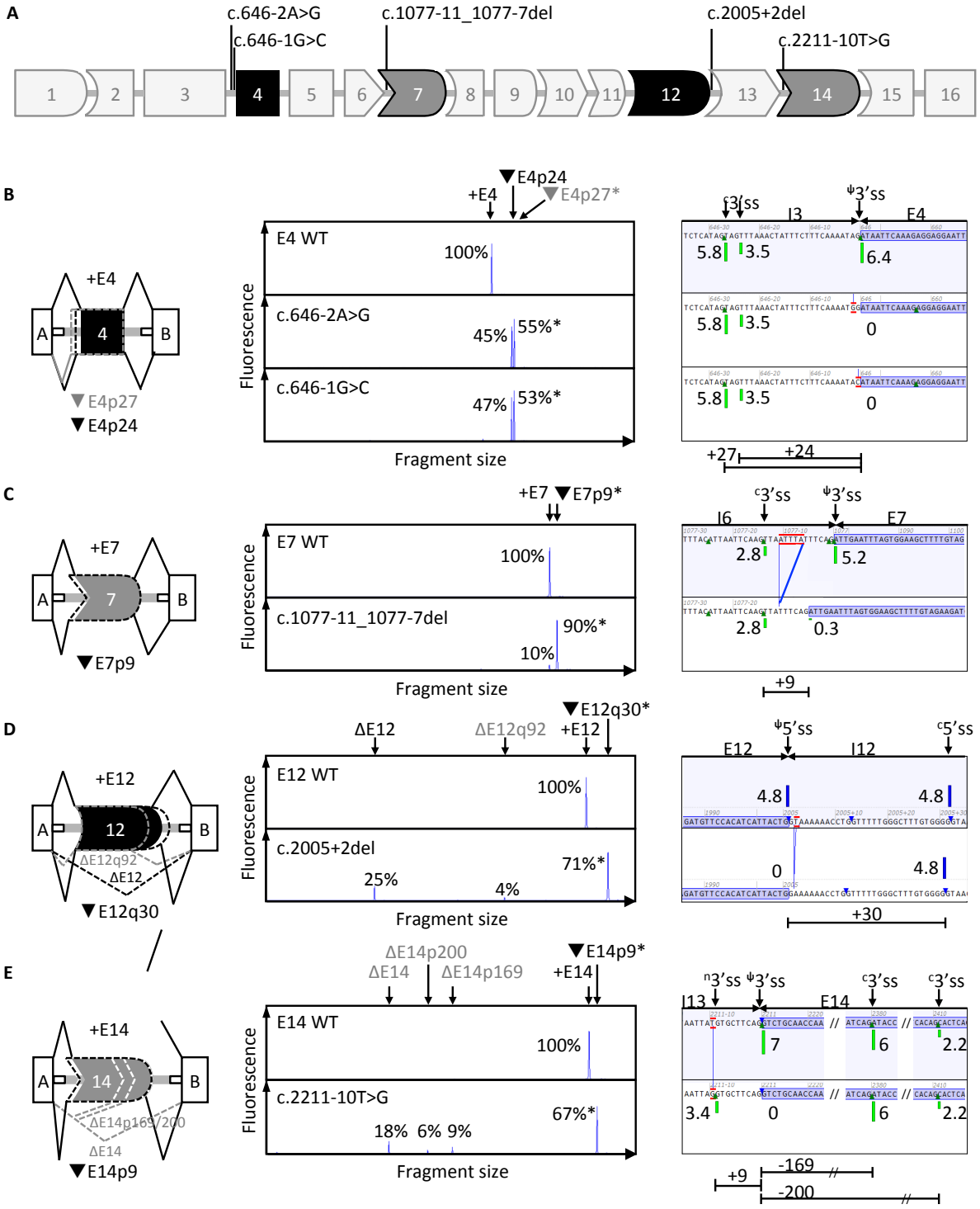

A
c.646-2A>G
c.646-1G>C
c.1077-11_1077-7del
c.2005+2del
c.2211-10T>G
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
B
+E4
A
4
B
E4p27
E4p24
E4p24
+E4
E4p27*
E4 WT
100%
c.646-2A>G
45%
55%*
c.646-1G>C
47%
53%*
Fluorescence
Fragment size
c3'ss
Ψ3'ss
I3
E4
5.8
3.5
6.4
5.8
3.5
0
5.8
3.5
0
+27
+24
C
+E7
A
7
B
E7p9
+E7
E7p9*
E7 WT
100%
c.1077-11_1077-7del
90%*
10%
Fluorescence
Fragment size
c3'ss
Ψ3'ss
I6
E7
2.8
5.2
2.8
0.3
+9
D
+E12
A
12
B
ΔE12q92
ΔE12
E12q30
E12q30*
ΔE12
ΔE12q92
+E12
E12 WT
100%
c.2005+2del
25%
4%
71%*
Fluorescence
Fragment size
Ψ5'ss
c5'ss
E12
I12
4.8
4.8
0
4.8
+30
E
+E14
A
14
B
ΔE14p169/200
ΔE14
E14p9
E14p9*
ΔE14p200
ΔE14
ΔE14p169
+E14
E14 WT
100%
c.2211-10T>G
18%
6%
9%
67%*
Fluorescence
Fragment size
n3'ss
Ψ3'ss
c3'ss
c3'ss
I13
E14
7
6
2.2
3.4
0
6
2.2
+9
-169
-200

**Figure 4**

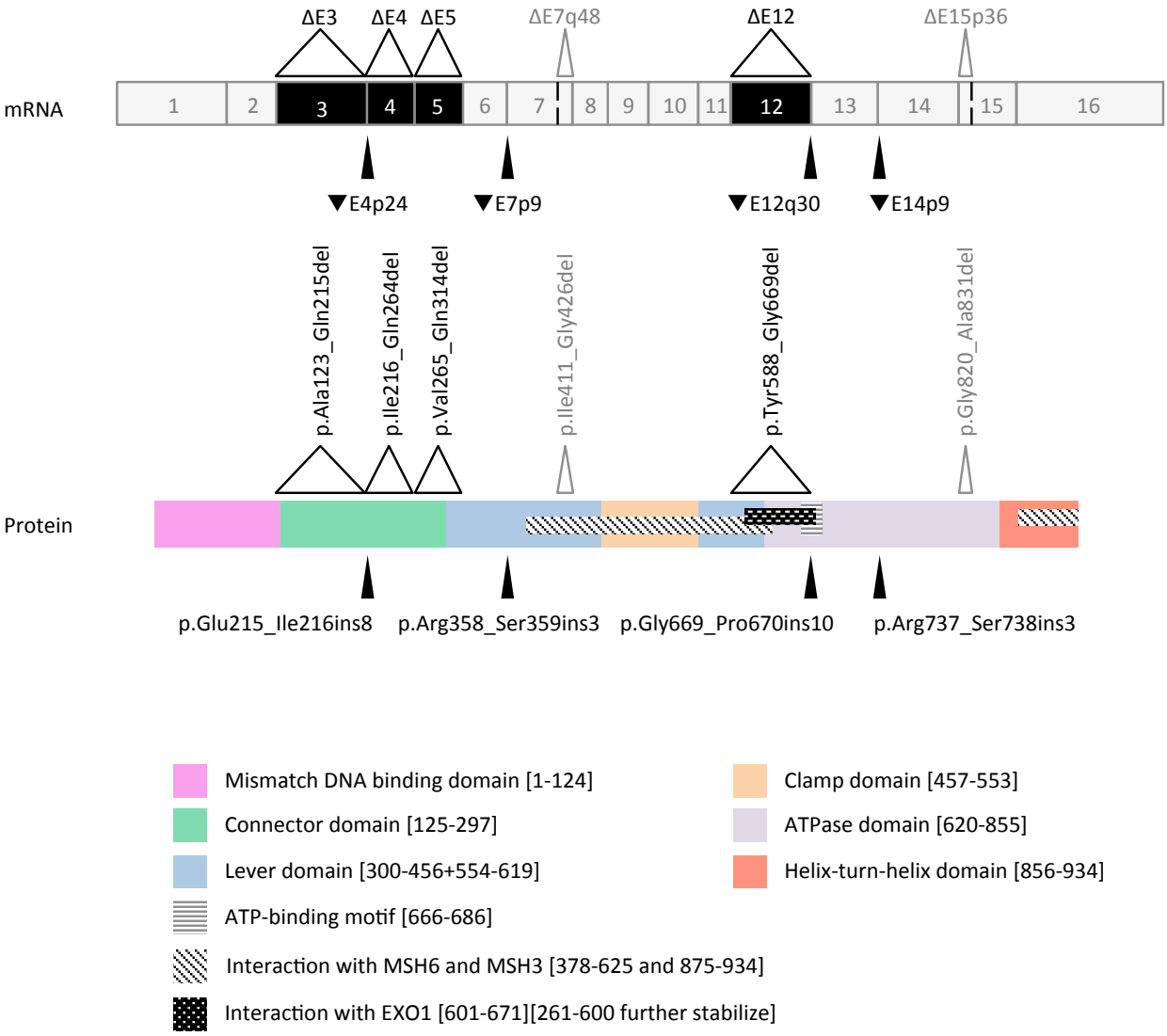
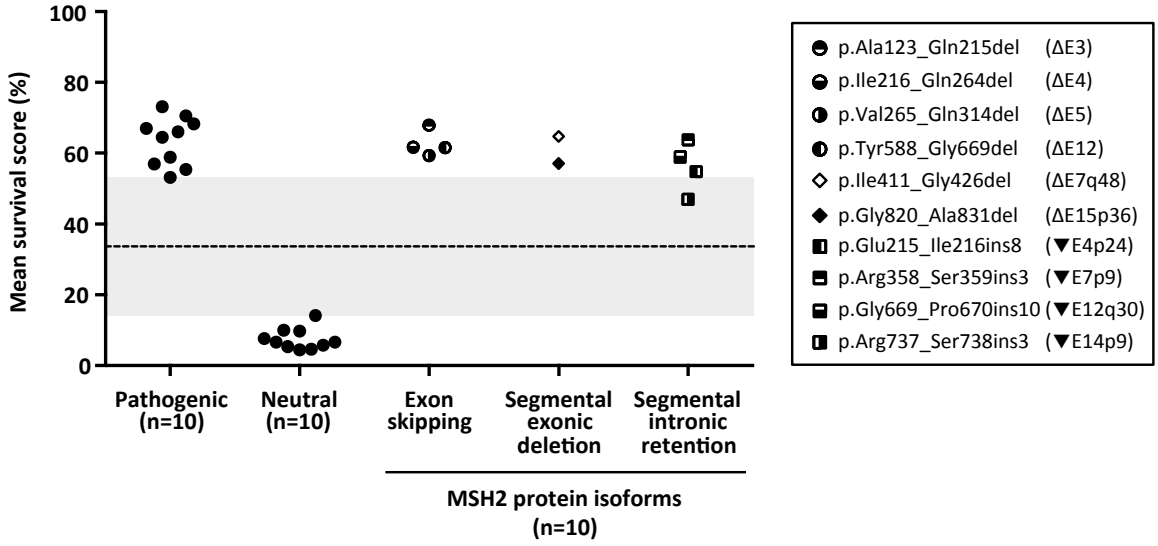

ΔE3
ΔE4
ΔE5
ΔE7q48
ΔE12
ΔE15p36
mRNA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
▼E4p24
▼E7p9
▼E12q30
▼E14p9
p.Ala123_Gln215del
p.Ile216_Gln264del
p.Val265_Gln314del
p.Ile411_Gly426del
p.Tyr588_Gly669del
p.Gly820_Ala831del
Protein
p.Glu215_Ile216ins8
p.Arg358_Ser359ins3
p.Gly669_Pro670ins10
p.Arg737_Ser738ins3
Mismatch DNA binding domain [1-124]
Connector domain [125-297]
Lever domain [300-456+554-619]
ATP-binding motif [666-686]
Interaction with MSH6 and MSH3 [378-625 and 875-934]
Interaction with EXO1 [601-671][261-600 further stabilize]
Clamp domain [457-553]
ATPase domain [620-855]
Helix-turn-helix domain [856-934]

**Figure 5**


Mean survival score (%)
100
80
60
40
20
0
Pathogenic (n=10)
Neutral (n=10)
Exon skipping
Segmental exonic deletion
Segmental intronic retention
MSH2 protein isoforms (n=10)
p.Ala123_Gln215del (ΔE3)
p.Ile216_Gln264del (ΔE4)
p.Val265_Gln314del (ΔE5)
p.Tyr588_Gly669del (ΔE12)
p.Ile411_Gly426del (ΔE7q48)
p.Gly820_Ala831del (ΔE15p36)
p.Glu215_Ile216ins8 (▼E4p24)
p.Arg358_Ser359ins3 (▼E7p9)
p.Gly669_Pro670ins10 (▼E12q30)
p.Arg737_Ser738ins3 (▼E14p9)