



HAL
open science

Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation

Vincent Jonchère, Hugo Montémont, Enora Le Scanf, Aurélie Siret, Quentin Letourneur, Emmanuel Tubacher, Christophe Battail, Assane Fall, Karim Labreche, Victor Renault, et al.

► To cite this version:

Vincent Jonchère, Hugo Montémont, Enora Le Scanf, Aurélie Siret, Quentin Letourneur, et al.. Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation. *Genome Biology*, 2024, 25, pp.210. 10.1186/s13059-024-03340-5 . hal-04909880

HAL Id: hal-04909880

<https://hal.science/hal-04909880v1>

Submitted on 24 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH

Open Access



Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation

Vincent Jonchère^{1†}, Hugo Montémont^{1†}, Enora Le Scanf^{2,3†}, Aurélie Siret¹, Quentin Letourneur¹, Emmanuel Tubacher⁴, Christophe Battail⁵, Assane Fall¹, Karim Labreche¹, Victor Renault⁴, Toky Ratovomanana¹, Olivier Buhard¹, Ariane Jolly⁶, Philippe Le Rouzic¹, Cody Feys¹, Emmanuelle Despras¹, Habib Zouali⁴, Rémy Nicolle⁷, Pascale Cervera^{1,8}, Magali Svrcek^{1,8}, Pierre Bourgoin^{1,8}, Hélène Blanché⁴, Anne Boland⁵, Jérémie Lefèvre^{1,9}, Yann Parc^{1,9}, Mehdi Touat^{1,10}, Franck Bielle¹¹, Danielle Arzur^{2,3}, Gwennina Cueff^{2,3}, Catherine Le Jossic-Corcós^{2,3}, Gaël Quéré^{2,3}, Gwendal Dujardin^{2,3}, Marc Blondel^{2,3}, Cédric Le Maréchal^{2,3}, Romain Cohen^{1,12}, Thierry André^{1,12}, Florence Coulet^{1,13}, Pierre de la Grange⁶, Aurélien de Reyniès⁷, Jean-François Fléjou^{1,8}, Florence Renaud¹, Agusti Alentorn¹, Laurent Corcos^{2,3†}, Jean-François Deleuze^{4,5†}, Ada Collura^{1†} and Alex Duval^{1,13*†} 

[†]Vincent Jonchère, Hugo Montémont, and Enora Le Scanf are co-first authors.

[†]Laurent Corcos, Jean-François Deleuze, Ada Collura, and Alex Duval are co-senior authors.

*Correspondence: alex.duval@inserm.fr

¹ Sorbonne Université, INSERM, Unité Mixte de Recherche Scientifique 938 and SIRIC CURAMUS, Centre de Recherche Saint-Antoine, Equipe Instabilité Des Microsatellites Et Cancer, Equipe Labellisée Par La Ligue Nationale Contre Le Cancer, 75012 Paris, France
Full list of author information is available at the end of the article

Abstract

Background: Microsatellite instability (MSI) due to mismatch repair deficiency (dMMR) is common in colorectal cancer (CRC). These cancers are associated with somatic coding events, but the noncoding pathophysiological impact of this genomic instability is yet poorly understood. Here, we perform an analysis of coding and noncoding MSI events at the different steps of colorectal tumorigenesis using whole exome sequencing and search for associated splicing events via RNA sequencing at the bulk-tumor and single-cell levels.

Results: Our results demonstrate that MSI leads to hundreds of noncoding DNA mutations, notably at polypyrimidine U2AF RNA-binding sites which are endowed with *cis*-activity in splicing, while higher frequency of exon skipping events are observed in the mRNAs of MSI compared to non-MSI CRC. At the DNA level, these noncoding MSI mutations occur very early prior to cell transformation in the dMMR colonic crypt, accounting for only a fraction of the exon skipping in MSI CRC. At the RNA level, the aberrant exon skipping signature is likely to impair colonic cell differentiation in MSI CRC affecting the expression of alternative exons encoding protein isoforms governing cell fate, while also targeting constitutive exons, making dMMR cells immunogenic in early stage before the onset of coding mutations. This signature is characterized by its similarity to the oncogenic U2AF1-S34F splicing mutation observed in several other non-MSI cancer.



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Conclusions: Overall, these findings provide evidence that a very early RNA splicing signature partly driven by MSI impairs cell differentiation and promotes MSI CRC initiation, far before coding mutations which accumulate later during MSI tumorigenesis.

Keywords: Colorectal cancer, Mismatch repair deficiency, Microsatellite instability, Whole-exome and RNA sequencing, Alternative splicing, Polypyrimidine U2AF binding site, Colonic crypt, Cancer initiation, Cell differentiation

Background

Genomic instability induces diversity that facilitates the development of the hallmark properties of tumor cells [1]. Following impairment of the DNA mismatch repair (MMR) system, high levels of genomic instability are observed in cancers with the mutator phenotype. This results in the accumulation of numerous mutations, particularly in repetitive DNA sequences such as microsatellites (MS) [2, 3]. This type of genomic instability is characteristic of the microsatellite instability (MSI) phenotype in some tumors [4–6]. Tumors presenting MSI can arise in patients with a hereditary cancer syndrome known as Lynch syndrome (LS) or can occur sporadically, as in approximately 15–20% of all colorectal, gastric, and endometrial cancers [7]. In mismatch repair-deficient (dMMR) tumor cells, it has been clearly demonstrated that the level of instability of a MS is positively and strongly correlated with its size [8, 9]. To date, somatic MS alterations leading to changes in gene homeostasis have been primarily observed in small coding MS-containing genes with low instability and in large microsatellites (i.e., > 10 bp) that are not present in the coding portion of the human genome due to the counterselection of such highly unstable sequences during the evolution of species (for review see [10]). Although some of these coding MS mutations have been proposed to play a role in tumorigenesis, there is no real functional evidence behind the claim in many cases; their occurrence has been mostly related to the synthesis of numerous aberrant immunogenic neoantigens [11] and the infiltration of MSI tumors with activated cytotoxic T cell lymphocytes (CTLs) and Th1 cells, thus creating a hostile, antitumor microenvironment. Consequently, MSI colorectal cancer (CRC) has a more favorable prognosis than microsatellite-stable (MSS) CRC, at least for localized tumors [12], and can be treated effectively with immune checkpoint inhibitors [13], although resistance to these new drugs can still develop [14–16]. In addition, DNA alterations due to MSI in cancer are much more common in the noncoding region of the genome [8, 9]. However, the functional consequences of MSI noncoding mutations remain largely unknown.

Human genes are generally composed of multiple (coding) exons interspersed with (noncoding) introns that undergo splicing to generate mature messenger RNAs (mRNAs) and subsequently generate proteins (reviewed in [17]). Pre-mRNA splicing is a complex process that involves constitutive intron removal and ligation of most exons in the gene as well as the inclusion or exclusion of certain exons from the final mRNA [17]. Alternative RNA splicing is observed at physiologically regulated exons in more than 95% of the 19,804 transcribed human coding genes [17]. This process enables the production of thousands of alternative mRNAs encoding numerous self-protein isoforms, i.e., proteins endogenously produced by DNA transcription and translation, generated specifically within a tissue during cell differentiation (as opposed to non-self-proteins that are not created within the body of the organism of interest, and subsequently can be

targeted and attacked by the immune system) [17, 18]. Alternative mRNA splicing is the main contributor of changes in the transcriptome as stem cells differentiate into tissue progenitor cells, notably in the colonic crypt [19]. Accumulating evidence indicates that both alternative and constitutive exons are aberrantly spliced during cancer development, in part due to somatic mutations resulting from genome instability [20]. Excessive alternative splicing in MSI CRCs compared to MSS CRCs has recently been observed using long-read nanopore sequencing, although the authors did not assess the reason for this observation or the pathophysiological roles that these alternative transcripts may play in these tumors [21]. Interestingly, numerous MS that can be highly unstable in the genome of MSI cancers due to their large size are located near the polypyrimidine tract (PY) of the intron/exon boundary [22, 23]. These noncoding MS are important *cis*-acting sequence elements for the binding of *trans*-acting core spliceosomal factors such as U2AF2 upstream of the intronic AG splice acceptor site which binds U2AF1 to guide intron removal during pre-mRNA splicing [24].

The MS distribution within human genes is not homogenous, with intronic regions displaying more frequent and larger MS than coding exons (see Fig. 1A left panel). A highly MS-enriched intronic region is often observed in the PY (Fig. 1A, right panel), and the PY is a major *cis*-acting determinant of pre-mRNA splicing [25] (Fig. 1B). In the present work, we hypothesized (WH1) that RNA changes specific to MSI CRCs occur because of the high mutational burden of these tumors, and these mutations could destabilize the physiological splicing process, which is dependent on the binding of *trans*-factor complexes from the spliceosome to the PY *cis*-sequence sites of pre-mRNAs (Fig. 1B and 1C). We also hypothesized (WH2) that MSI-related mutations disrupt splicing early during tumor evolution, as these changes affect to large microsatellites, which have a very high mutation rate in dMMR cells, and that these alterations thus facilitate MSI tumor initiation (Fig. 1B and 1C). Finally, we assumed (WH3) that the splicing changes due to MSI should mimic those observed in other tumors with disruption of U2AFs complexes, notably composed of U2AF1 and U2AF2, since these spliceosomal factors normally bind RNA at the PY for correct splicing (Fig. 1B and 1C). To assess these hypotheses, whole exome sequencing was performed to study MSI at all the human intron/exon boundaries and up to 50 bp into the intronic sequence. We analyzed a series of dMMR precancerous lesions, i.e., a pool of 10 different dMMR crypt foci from 8 MSI CRC patient and 14 dMMR adenoma patients, with the aim of examining the role of MS in splicing in the very early steps of the MSI-driven tumorigenic process (Fig. 1D). In addition, 46 MSI CRC samples with matched normal mucosa samples were also analyzed to investigate the mutational burden at the same noncoding MS in established colorectal tumors. RNA sequencing was performed on 101 established MSI CRC specimens including 46 MSI CRC specimens with whole-exome sequencing data, 32 MSS CRC specimens, and matched normal mucosa samples ($n = 133$). Due to the lack of sufficient material in terms of quantity and quality, we were unable to sequence the RNA from the microdissected dMMR crypts and adenoma. Finally, we examined alternative splicing in CRC in relation to the MSI or MSS phenotype at the single-cell (SC) level using a public dataset from patients with MSI (34 patients—33,333 cells) or MSS (28 patients—33,334 cells) CRC [26] (Fig. 1D). Overall, our results provide evidence that a new, very early RNA splicing pathway acts as a mimic of the mutation in *trans* of U2AF1

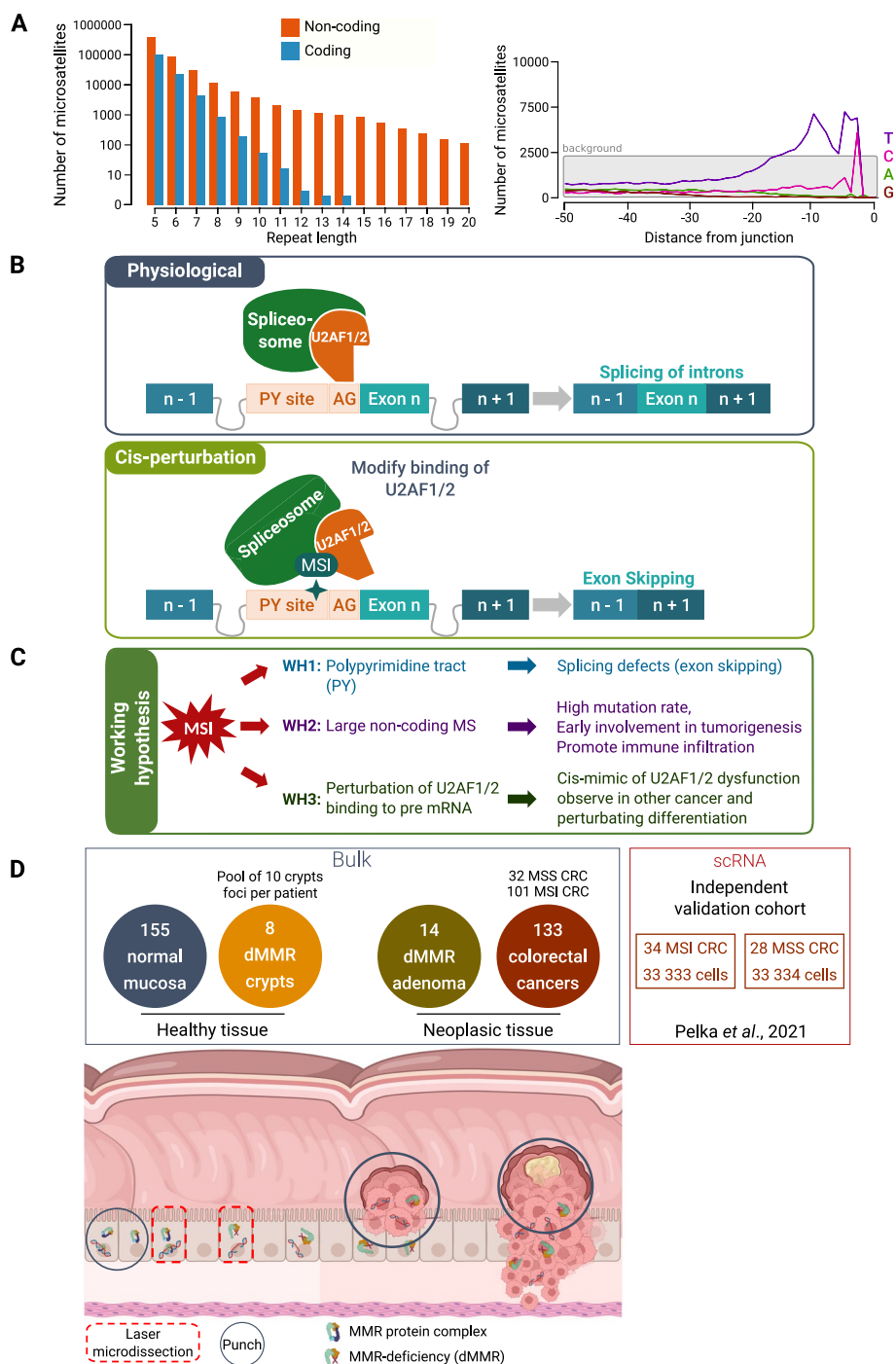


Fig. 1 Frequently mutated intronic microsatellites may cause aberrant exon skipping in microsatellite instability (MSI) colorectal cancer (CRC). **A** Left panel, genomic distribution of microsatellite (MS) mutations (log10 scale) in two gene regions (exonic and intronic) according to repeat length. Right panel, distribution of all flanking MS (3' splice acceptor site, 3' FMS) according to their distance from the intron–exon boundary. The distribution of these MS signals is indicated according to their nucleotide composition (violet: thymine, pink: cytosine, green: adenosine, brown: guanine). The background corresponds to the distribution of all MS across the human genome. **B** Upper panel, schematic representation of physiological splicing. PY, polypyrimidine tract; AG, acceptor site. Lower panel, schematic representation of MSI *cis*-perturbations of splicing. **C** Working hypothesis (WH). **D** Experimental design to investigate splicing *cis*-perturbations in MSI CRC specimens

in several other cancers. They also establish how this pathway, which is partly driven by MSI at the PY, is likely to promote dMMR cell transformation and MSI CRC initiation by impairing colonic cell differentiation, far before coding MSI mutations which accumulate later in established MSI CRC.

Results

***Cis*-splicing intronic polypyrimidine tract regions are enriched in DNA microsatellites frequently mutated very early in dMMR colonic crypts and colonic dysplasia prior to their transformation into MSI CRCs**

In Fig. 2A, we outlined our approach for longitudinal analysis of MSI at crucial steps during colorectal tumorigenesis, i.e., from MMR crypt to adenoma to established adenocarcinoma. Through quantitative analysis of MSI in the genome using MSICare, MSI was already observed in dMMR crypts, albeit at moderate levels, and rapidly reached a plateau early in adenoma that was sustained adenocarcinoma (Fig. 2B and also Additional file 1: Fig. S1A). Due to differences in microsatellite size, we observed a much greater number of unstable noncoding MS in the vicinity of the PY than unstable coding MS (Fig. 2C right panel). In a dynamic way, our data establish that the level of MSI at noncoding MS whose sizes are frequently greater than 10 nucleotides, reaching 15–20 base pairs or more, is already very high in the dMMR crypts, whereas the instability of coding MS whose lengths exceptionally exceed 10 nucleotides in the human genome is quantitatively much lower (Fig. 2C left panel and Additional file 1: Fig. S1B). Among the dMMR cryptic lesions we were able to examine, the level of MSI gradually increased from monocryptic dMMR foci to slightly more advanced oligo/polycryptic dMMR foci (Fig. 2D). It is clear that in the dMMR crypts, it is the instability of MS very close to the AG site that already prevails within the 3' portion of PY known to play a key role in splicing, with a further increase in the number of such MSI events during the progression towards cancer (Fig. 2E).

Massive perturbation of splicing in MSI CRC

We then investigated pre-mRNA splicing and its disruption in MSI CRC by performing RNA sequencing of the bulk tumor. We deliberately limited our analyses to exon insertion/exclusion within mRNAs as exon insertion/exclusion represents the type of RNA alteration that is most likely to be directly influenced by *cis*-acting elements related to MSI at intron boundaries, notably at the PY. The inclusion or exclusion of all human 3' exon-flanking a microsatellite (3' Exon-FMS) was measured by the percent spliced-in index (PSI) as previously described [28]. This is a standard method that represents the ratio between reads that include or exclude specific exons (Additional file 1: Fig. S2). Overall, we identified 985 exons that exhibited significantly altered expression in MSI CRC vs MSS CRC and normal colon tissues (Fig. 3A). These included 839 deregulated skipping events at the 3' Exon-FMS in MSI CRC tissue compared to normal colonic tissue, and 276 additional events in MSI CRC tissue compared to MSS CRC tissue, of which 146 were new exons to be incorporated into the signature. A greater frequency of deregulated exon splicing events was observed in MSI CRC compared to MSS CRC (two-sided p -value $< 2 \times 10^{-20}$) (Additional file 1: Fig. S3 and Additional file 2: Table S1). Most splicing changes involved exon exclusion (see also the data on aberrant exon

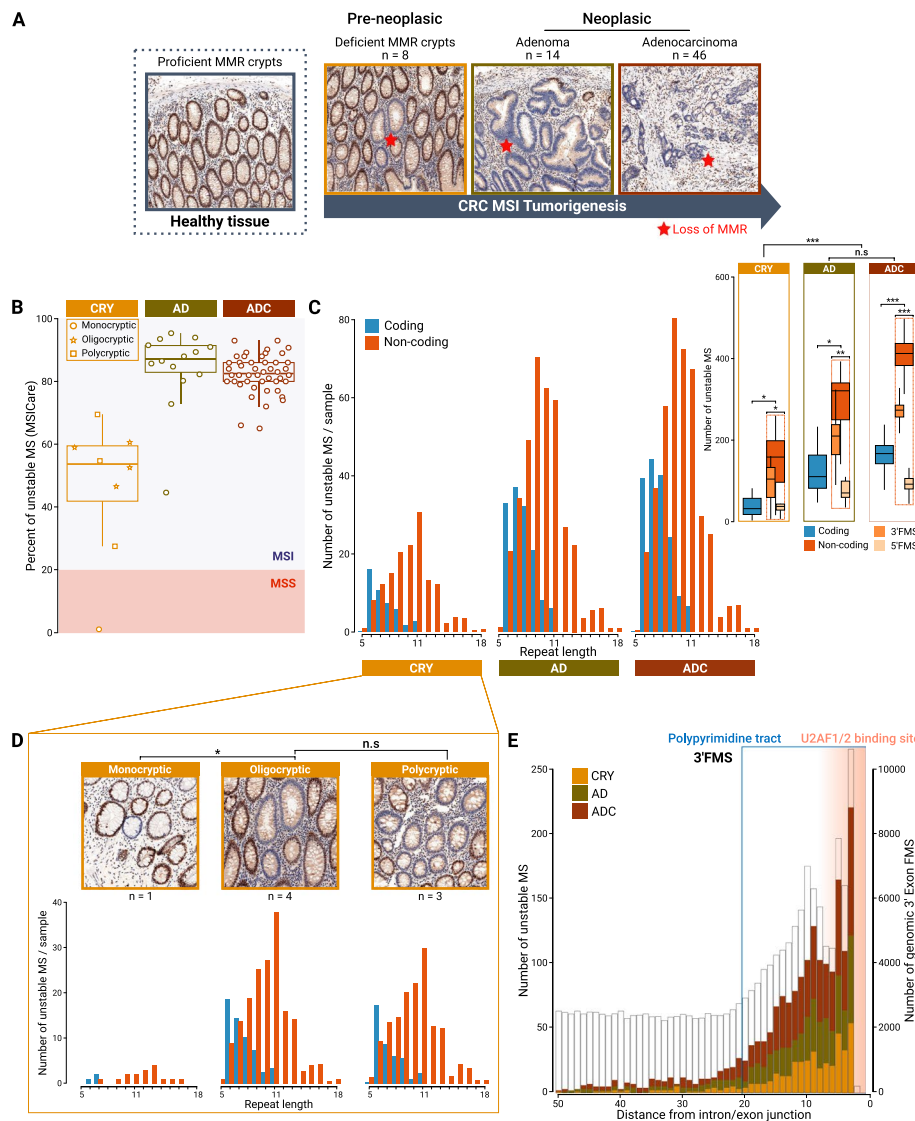


Fig. 2 MSI-driven mutational events in non-coding regions lead mRNA changes which appear in the early stages of MSI tumorigenesis. **A** Histological description of MSI CRC tumorigenesis, from preneoplastic lesions (dMMR crypts) to neoplasms of different severity (adenoma and adenocarcinoma). Immunohistochemical staining of the MSH2 protein was performed. **B** The MSI status for each sample was defined by MSICare [27]. **C** Left panel, number of unstable MS by sample according to their length and considering their genomic position (coding or non-coding). **C** Right panel, boxplot representing the total number of unstable MS according to their region and condition. Statistics: Student's *t*-test, **P*-value < 0.05, ***P*-value < 1.10⁻³, ****P*-value < 1.10⁻⁴, *****P*-value < 1.10⁻⁵. **D** Loss of MMR in the colonic crypt was categorized into three groups, depending on the number of dMMR crypts in the foci: monocryptic (1), oligocryptic (2–4), and polycryptic (>4). In all the groups, crypts are identical and morphologically healthy, **P*-value < 0.05. **E** Distribution of 3' FMS according to the distance of the 3' splice site (AG); the genomic background distribution of all flanking MS (*n* = 335,564) is indicated (phantom white bars)

inclusion in MSI CRC in Additional file 1: Fig. S3). These changes were detected at high levels in all MSI CRC samples and were a distinctive feature of the transcriptome profile of both hereditary (Lynch syndrome-related) and sporadic MSI CRC, independent of TNM stage (stage 2 or 3) or the presence of *BRAF/KRAS* driver mutations. When we explored possible association of each exon skipping event from this splicing signature

with clinical (i.e., TNM staging, sex, Lynch vs. sporadic inferred status, location of the tumor in the colon) or molecular variables (i.e., *KRAS* or *BRAF* mutations, type of MMR defects, *MLH1* promoter methylation status), none of the tested associations was significant after adjusting *p*-values (data not shown). The average number of skipping events per tumor was similar to the number of coding MS mutations in this MSI CRC cohort (Fig. 3B) [8].

The 985 deregulated exons were then characterized as products of alternative or constitutive splicing. The expression patterns of the exons were first analyzed in 133 normal colorectal mucosa samples (57,030 analyzed exons). As expected, the majority of these exons (88.6%) were constitutively expressed. The remaining exons (11.4%) were alternatively expressed [17, 18] (Fig. 3C left panel). In MSI CRC samples, a minority of the deregulated exons were constitutive ($n = 293/985$, 29.7%), whereas the majority ($n = 692/985$, 70.3%) belonged to the 11.4% of exons that displayed an alternative expression pattern in normal colonic mucosa (significant enrichment by 6.2-fold; two-sided p -value $< 2 \times 10^{-16}$) (Fig. 3C right panel, Additional file 1: Fig. S4). These alternative exons were deregulated at low, moderate, or high frequencies in MSI CRC (Fig. 3D left panel), as opposed to the constitutive exons, which were mostly skipped at lower frequencies (Fig. 3D right panel) (two-sided p -value $< 2 \times 10^{-16}$). Finally, we confirmed that the frequency of exon skipping was also high among the above 839 candidate exons in independent cohorts of patients with MSI tumors from the TCGA. These results were obtained considering not only MSI CRC but also MSI gastric and endometrial cancer, highlighting that the splicing signature is a shared pathophysiological feature of these frequent MSI tumors derived from 3 distinct primary sites and is most likely a common feature across MSI tumors (Fig. 3E).

(See figure on next page.)

Fig. 3 MSI-driven mutational events in non-coding regions lead to the skipping of mainly alternatively regulated exons in MSI CRC. **A** Left panel, Venn diagram showing the overlap between exon skipping events in MSI CRC tissue compared to normal colonic tissue and to MSS CRC tissue. **B** Bar plot displaying the number of mutations in coding MS (expected frameshift mutations) and the number of skipped exons across the whole exome for 46 non-metastatic MSI tumors (TNM stage 2 or 3). **C** Left panel, pie chart representing the portion of alternative and constitutive exons in normal colorectal tissue ($n = 57,030$ exons, $n = 133$ normal tissue samples). Right panel, pie chart presenting the portion of the same alternative and constitutive exons whose expression was deregulated in MSI tumors. The results of the chi-square test between the two pie charts are indicated. **D** Distribution of exon skipping events according to their frequency in MSI tumors. Left and right panels, exons classified as constitutive ($n = 293$) or alternative ($n = 692$), respectively, in normal colonic tissue. Bottom panel, example of IGV-Sashimi plots showing the read coverage for 2 examples of significantly deregulated exons between tumors and normal tissue (right, constitutive and left, alternative). **E** Percentage overlap of MSI exon skipping between our study cohort and three independent TCGA cohorts. The number of overlapping events and total number of events analyzed are indicated in brackets. The *P*-values of the enrichments are indicated. **F** Cell distribution in MSS (Top panel) or in MSI (Bottom panel) according to sample status (tumor or normal tissue) (Left) and cell subset (Right). The total number of cells used was the same that in the Pelka et al. [26] study ($n = 371,223$) (more details in Fig. S5, also showing the distribution of the different cell types according to the MSI/MSS status of tumors). However, when we performed the splicing analysis, the total number of cells available was reduced to $n = 128,370$, but with the same distribution of cell types, i.e., no statistically significant difference was observed in the distribution of cell types in MSI/MSS tumor samples in the population and in the dataset with PSI data. The different number of cells were subsampled, and we performed a bootstrapping analysis (10,000 permutations), considering the same distribution of PSI values according to TA or colonocytes in MSS/MSI. In both situations, more than 100 cells in each condition, allowed to identify statistically significant differences. **G** Boxplot of quantitative index of alternative splicing events in transit amplifying cells (TA) according to MSI tumor status

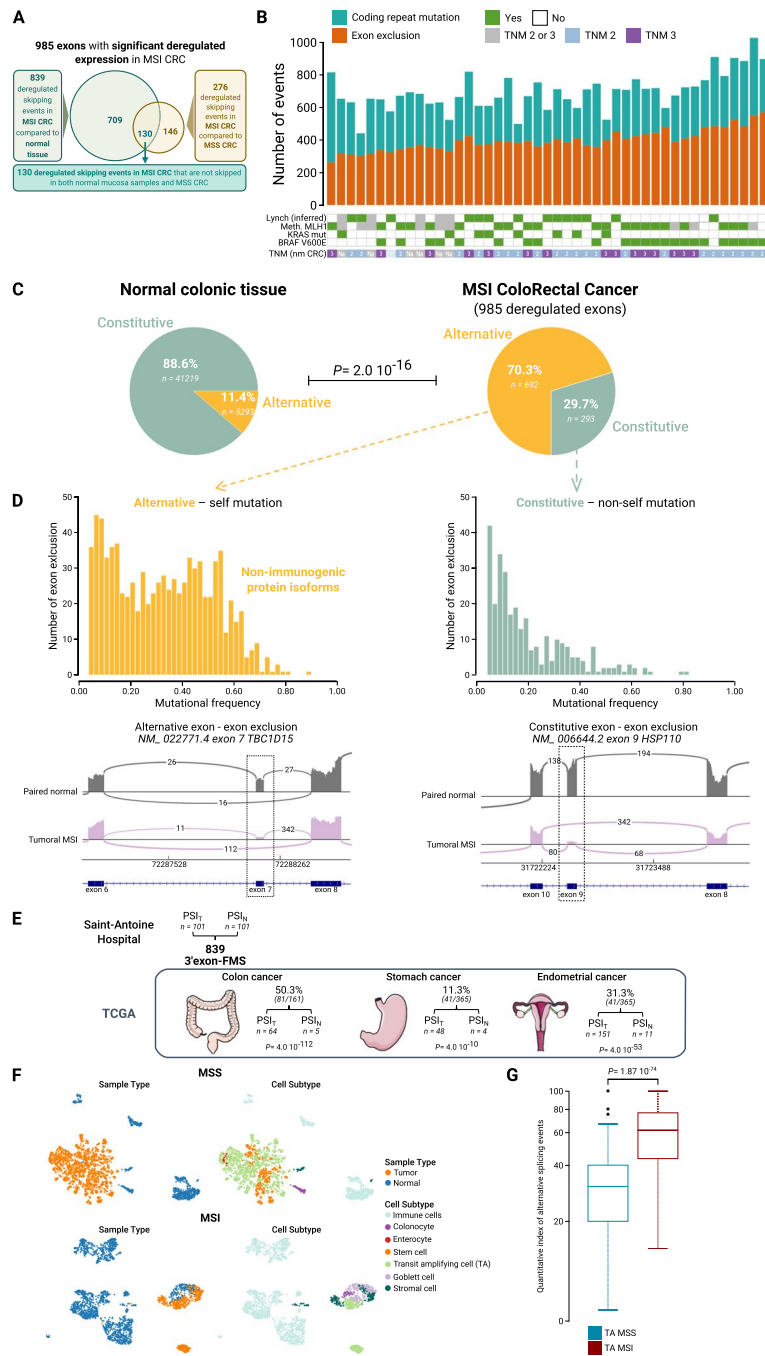


Fig. 3 (See legend on previous page.)

Confirmation of the increased prevalence of alternative splicing events in MSI versus MSS CRC at the single-cell level

In line with the above results, we wanted to further investigate alternative splicing in CRC in relation to the MSI or MSS phenotype at the single-cell (SC) level. Independent public scRNA sequencing datasets from both MSI and MSS CRC were obtained from the Pelka et al. study [26]. In both types of CRC, i.e., MSI and MSS tumors, the predominance of tumor clones with the same TA (transit amplifying cells) state was

observed using Census (<https://doi.org/https://doi.org/10.1101/2022.10.19.512926>) [29] and CellTypist [30] (Fig. 3F), and the scRNA sequencing analyses therefore focused on these predominant clones within the bulk tumor. With the MARVEL package [31] which enables systematic and integrated splicing and gene expression analysis of single cells to characterize the splicing landscape and reveal biological insights, we were able to measure the degree of alternative exon inclusion from scRNA sequencing data in both MSI and MSS contexts. The overall number of alternative splicing events between transit-amplifying (TA) MSS and TA MSI cell types is shown in Fig. 3G. The results clearly show an overall highly significant increase in the number of such events in TA MSI clones compared to TA MSS malignant clones (p -value $< 2 \times 10^{-74}$). Notably, these analyses cumulatively detected all types of alternative splicing events and not only exon skipping. These findings support a strong link between MSI and alternative splicing during transcription in CRC at the single-cell level (see also Additional file 1: Fig. S5 for further details on the annotation used and the relationship of our annotation with the one used by Pelka et al. [26]).

Eon skipping events in MSI CRCs are partly due to MSI-related somatic deletions in intronic MS at the PY

We then sought to address whether a substantial proportion of splicing events in MSI CRCs are directly induced by MSI through a *cis*-acting mechanism at the PY intronic site. This was done by assessing whether somatic deletions in noncoding 3' exon-flanking MS would promote exon skipping in MSI tumors. We examined the status of DNA MS via whole-exome sequencing (up to 50 bp into the intronic sequence) in the initial MSI CRC cohort ($n=46$) and compared this to paired normal mucosa ($n=46$). Of the 985 deregulated exons, 277 MS/exon pairs were selected for this analysis because they showed sufficiently high coverage and at least one mutation at 3' noncoding MS in at least one MSI CRC sample to allow association studies (Additional file 3: Table S2). For the remaining 708 MS/Exon couples, a correlation could not be calculated because of the absence of somatic mutations affecting MS at the PY in MSI CRCs, and we assumed that the expression of these exons was not due to a direct mechanism in *Cis* at the PY.

In Fig. 4A, we show examples of genes in which the expression level of a single exon is closely related to the status of the DNA microsatellite in the PY. Overall, a significant correlation between somatic deletions at flanking 3' noncoding MS and exon skipping (PSI Index) was observed for 96 of these 277 candidates (34.7%) (Fig. 4B, left panel). As the size of somatic deletions in the intronic MS of MSI tumors became larger, the exclusion of these 96 exons in tumor RNAs increased. Alternative exons were more readily skipped by such cumulative somatic deletions due to MSI compared to constitutive exons in MMR-deficient tumors (Fig. 4B left panel, Additional file 1: Fig. S6, S7 and Additional file 4: Table S3). We further examined the impact of each of these 96 MSI-driven exon skipping on the overall expression of corresponding mRNAs or proteins in MSI CRC using an independent public CPTAC dataset (https://pdc.cancer.gov/pdc/browse/filters/pdc_study_id:PDC000109%7CPDC000116%7CPDC000117) that also included protein-level data. These data are shown for illustrative purposes only, being limited to the analysis of a much smaller number of patients with MSI CRC, without reaching significance in a number of cases, expectedly (Additional file 1: Fig. S8A).

Moreover, the majority of exons deregulated in MSI CRC are alternative exons, so that it is not expected that their skipping would have an impact on the overall expression of corresponding mRNAs, these events being more likely to affect the fine tuning of specific alternatively spliced mRNAs and corresponding protein isoforms in MSI tumor cells. By contrast, it also appears that certain microsatellites are unlikely to trigger exon skipping, even when they are frequently unstable in MSI CRC (Additional file 1: Fig. S8B and Additional file 3: Table S2). Compared to others, non-coding MS ($n = 96$) that were shown to directly influence RNA splicing in MSI CRC were primarily located very close to the intron/exon junction, within the functional polypyrimidine tract, just upstream of the U2AF1 (AG) splice acceptor 3' site [32] (Fig. 4B, right panel).

Functional analysis of the impact of MSI on gene splicing in a selected number of MS at the PY using gastrointestinal cancer cell lines

We next examined the impact of MSI on gene splicing by RTPCR in 10 gastrointestinal cancer cell lines, i.e., 5 MSI (SNU-1, Co115, HCT116, HCT8, LoVo) compared to 5 MSS cellular models used as controls (AGS, HGT1, HT29, N87, SW480). This was done by analyzing the effect of endogenous deletions at the RNA level in a small subset of noncoding MS from the ES96 list (Additional file 1: Fig. S9A), i.e., in *MRE11*, *HSP110* (also called *HSPH1*), *KDM6A*, and *TRAF3IP1* genes. Besides, we also investigated noncoding MS in *ATM*, *DNAJC18*, and *STAD-US* outside the list of mis-spliced genes provided in this study, because previous data from the literature (*ATM*) or online

(See figure on next page.)

Fig. 4 Splicing anomalies found in MSI CRCs are frequently caused by MSI-related mutations that occur very early in cancer development, as early as the untransformed dMMR crypt state. **A** Four examples of genes in which exon skipping is closely related to the status of the DNA MS in the PY at the intronic boundary. **B** Left panel, boxplots representing the percent spliced included (PSI) values according to deletion size for the 96 MS (ES96). The results of two-sided ANOVA between exons classified as alternative and those classified as constitutive were as follows: *** P -value $< .001$. In the background, each MS displaying a significant correlation is indicated by a line and dots. Right panel, distribution of 96 MS (ES96) at the intronic boundary (50 to 0 of intron/exon junction), showing an enriched localization very close to the U2AF1/2 binding region in the PY (AG site). **C** In vitro analysis of HSP110 expression. Upper panel, the RT-PCR products of HSP110 show cDNA fragments with a completed or partial skipped exon (HSP110DE9), in mutated MSI cell lines with large deletion (CO115) and small deletion (HCT116) respectively. Bottom panel: western blotting analysis of HSP110wt and HSP110DE9 mutant proteins in the 3 cell lines (CO115, HCT116 and HCT8). NSB*, non-specific band; MW, molecular weight. Uncropped images are available in additional file 3. **D** In vitro analysis of TRAF3IP1 gene. Upper panel, RT-PCR products of TRAF3IP1 showing cDNA fragments with a skipped exon (TRAF3IP1 DE6), in MSI cell lines (CO115 and HCT116). Middle panel, aberrant splicing event is also demonstrated by RT-PCR in endogenous and mutant (intronic MS mutation in the PY) TRAF3IP1 transcripts in both MSI and MSS transfected cells with minigene construction. EV, empty vector; WT, wild type; Splice Ratio = ratio of the intensity of the exon-lacking cDNA fragment to the intensity of the sum of exon-containing + exon lacking cDNA fragments. Lower panel, gel mobility shift assay. Nuclear protein extracts from MSS (left panel) and/or MSS (right panel) cells were incubated with TRAF3IP1 wild type or mutant labeled RNA probe. Shortened RNA probes do no longer allow the formation of large RNA/protein complexes. Uncropped images are available in Additional file 3. **E** Left panel, bar plots of microsatellites in the ES96 and C129 sub-signatures (consisting of coding MS covered in at least 50% of the samples and mutated in 30% of all lesions, $n = 129$) according to their mutational frequency in the early stage, i.e., dMMR crypts. The top 25% most frequently mutated MS (ES96/Coding) corresponded to MS in the first quartile when they were ranked by mutational frequency. Coding MS known to be highly mutated in MSI colorectal cancer are write in black. ES96 signature is significantly enriched in the top genes; * P -value $< 1.10^{-2}$. Right panel, non-exhaustive list of the most frequently mutated MS (ES96/Coding) in the early stage of MSI CRC tumorigenesis (more details in Additional file 7: Table S6). MS are order by mutational frequency in dMMR crypts

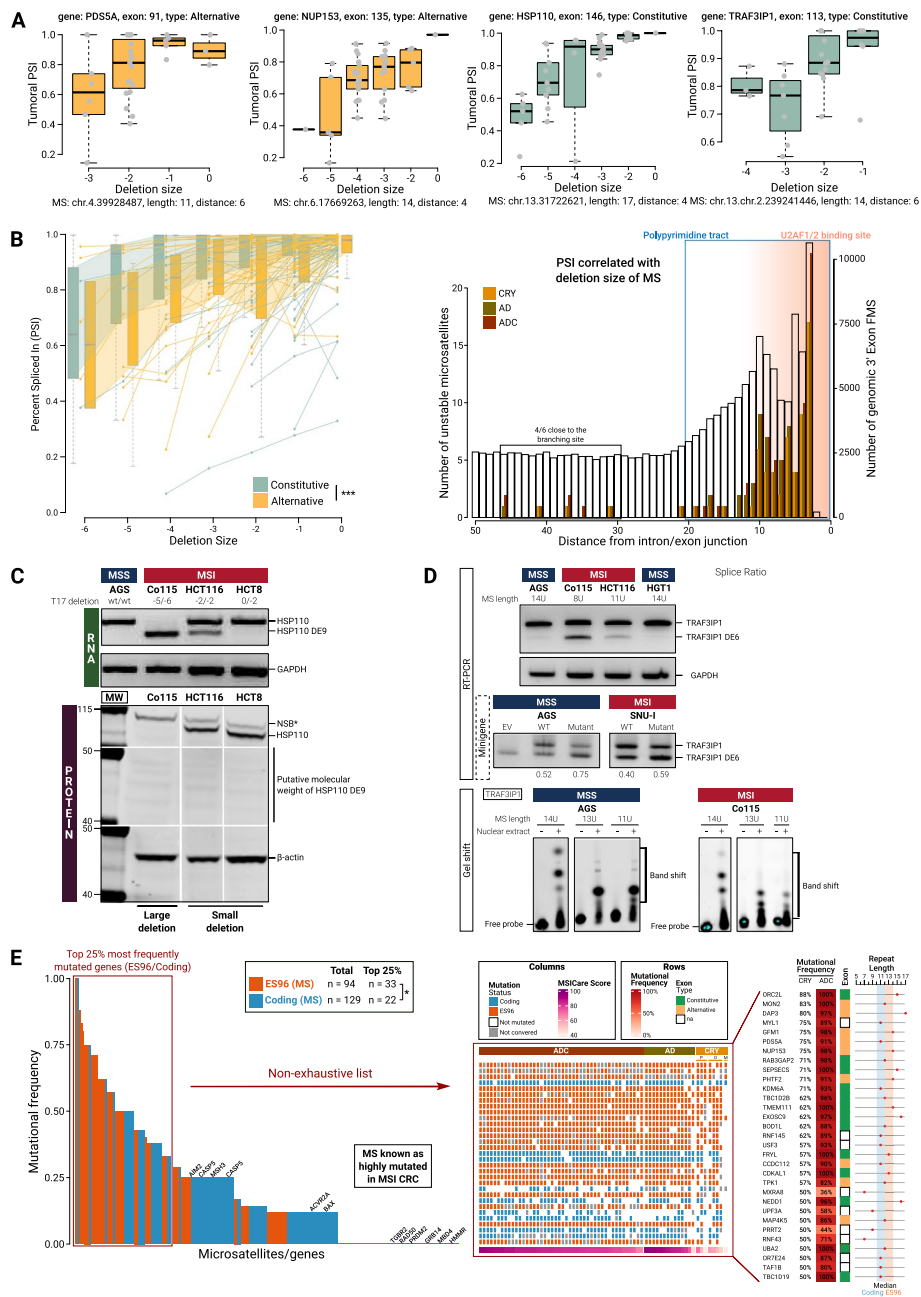


Fig. 4 (See legend on previous page.)

(*DNAJC18*, *STAD-US*) suggested a process of exon instability due to MSI at the PY in non-colorectal MSI tumors. At the protein level, as an example of a gene from the ES96 listing whose MSI-driven exon skipping leads to protein expression changes, we have further analyzed the case of the previously reported HSP110DE9 splicing mutation [22]. In brief, for this candidate gene previously shown to be sensitive to NMD in MSI CRC [33], we show that large deletions of the intronic MS at the PY lead to exon 9 skipping in the RNA and extinction of wild-type HSP110 protein expression. The HSP110DE9 mutant protein is not detected due to degradation of the corresponding mRNA during translation (Fig. 4C). We next selected PY nucleotide MS deletions from single introns of

the *DNAJC18*, *PTP4A2*, and *TRAF3IP1* genes (with 8, 14, and 14 consecutive T, respectively) that are frequently mutated in digestive cancers. These introns, harboring 1 and/or 3 T-deletions in the PY, showed increased level of exclusion of the downstream exon in mRNA of mutated MSI cancer cells (Co115 and HCT116 lines) but not in that of MSS AGS or HGT-1 cancer cells (Fig. 4D and Additional file 1: Fig. S9B). Importantly, these exon skipping events were replicated in minigene transfection experiments [34] regardless of the MSI or MSS status of the recipient cells, suggesting that such exon exclusions were essentially governed in these models by *cis*-microsatellite elements but not by *trans*-acting factors (Fig. 4D and Additional file 1: Fig. S9C).

Furthermore, to determine if RNA probes/protein interactions could be evidenced *in vitro* and modified due to MSI at the PY, we performed gel retardation assays with nuclear protein extracts from both MSI and MSS cells, using various PY as baits. This showed the occurrence of several complexes, in both MSI and MSS tumor cells, with similar migration patterns. Strikingly, the larger complexes disappeared when using probes deleted of 1 (DNAJC18) or 3 U (PTP4A2 and TRAF3IP1), suggesting that spliceosomal subunits could no longer form and/or assemble properly at the intron–exon junction when the PY is mutated (Fig. 4D and Additional file 1: Fig. S9D). In addition, in minigene transfection experiments, silencing U2AF1 gene showed increased skipping of the tested exons (increased Splice Ratios), indicating that U2AF1 is decisive in the control of spacing activity for these 3 minigene substrates (Additional file 1: Fig. S10A). Finally, to determine if U2AF2, the dimerization partner of UAF1 binding RNA at the PY, was able to interact *in vivo* within the complexes from wild-type or mutant PY, we performed RNA immunoprecipitation experiments. Interaction of U2AF2 with nuclear protein(s) was clearly observed in both MSS and MSI tumor cells *in vivo*, but we saw no difference between them, showing that some additional determining factors (proteins and/or RNAs) could be essential for splice site recognition in live cells (Additional file 1: Fig. S10B).

Some MSI-driven splicing mutations found in MSI CRCs sometimes occur very early during cancer development

Next, we sought to clarify the kinetics of occurrence of the above MSI-driven splicing mutations during the multistep process that facilitates cancer development in the colorectal mucosa. Among the 96 mutations associated with exon skipping in tumors, a majority, i.e., 61/96 (64.9%), were already present in dMMR colonic crypts, and a number of them were present in high frequencies (e.g., in *ORC2L* (88%), *PDS5A* and *NUP153* (75%), and *RAB3GAP2* (71%)) (Fig. 4E, Additional file 1: Fig. S11 and Additional file 5: Table S4).

Taken together, these results illustrate the earliness of many aberrant MSI-driven splicing MSI mutations in the dMMR crypt, prior to its transformation, in a much more prevalent way than what can be observed at the level of coding microsatellites at this pre-tumoral stage.

The MSI splicing signature primarily deregulates alternative mRNA isoforms that normally drive cell differentiation in the colonic crypt

We next sought to assess the functional relevance of the splicing signature identified here in CRC ($n=985$ exons). When we compared our MSI splicing signature ($n=985$

exons) to that of Habowski et al., who reported alternatively spliced mRNAs that drive cell differentiation in the normal murine colonic crypt [19], an enrichment in genes whose alternative splicing is modified as stem cells differentiate first into progenitors and then into mature cell types in the normal colonic crypt was observed [19] (Fig. 5A and Additional file 6: Table S5). This enrichment analysis revealed a total of 134 unique genes (DD134 gene signature). Compared to coding MS mutations in CRC, MSI-driven exon skipping events mainly deregulate the fine tuning of self-protein isoforms that are normally tightly regulated in the normal colonic mucosa under physiological conditions (Fig. 5B). Deregulation of these self-protein isoforms, while being expected to be well tolerated from an immune point of view, is likely to impair the differentiation of MMR-deficient cancer cells, in line with the poorly differentiated histological pattern that is a characteristic of MSI CRC as compared to MSS CRC, which is observed in the great majority of MSI tumors [7, 35]. In line with this, the expression of alternative exons in MSI CRC was significantly associated with several cancer-related pathways (Fig. 5C, left panel). Among them, some pathways are known to play an important role in tissue differentiation in the colon along the colonic crypt and also during colorectal cancer initiation, e.g., BMP/TGFBeta, Wnt/Wingless, and Hippo signaling pathways (two-sided p -value < 0.01) [36] (Fig. 5C, right panel). When investigating more particularly the pathways associated with the reduced DD134 exon signature, the association of this reduced signature with TGFBeta signaling remained significant (Fig. 5B, C and Additional file 1: Fig. S12).

Finally, to further illustrate that some noncoding MSI splicing mutations could affect the function of alternative protein isoforms, we carried out an *in silico* analysis of 5 of the 11 genes shared by the ES96 and DD134 signatures whose exon skipping due to MSI is an alternative exon. These genes are *AMBRA*, *DHX30*, *GOLGA4*, *NUP54* and *PAPOLA*. For 3 out of these 5 genes, it is highly probable that the loss of the alternative exon following MSI induces a functional change in the protein (i.e., in *DHX30*, *NUP54*, and *PAPOLA*), or even a loss of function due to potential exclusion of key functional domains (see further details in Additional file 1: Fig. S13).

(See figure on next page.)

Fig. 5 Changes in mRNA impair cell differentiation and mimic oncogenic U2AF1 inactivation in MSI CRC. **A** Schematic representation of the normal colonic crypt. **B** A total of 134 genes that overlapped between our MSI splicing signature and that of Habowski et al. that regroup alternative mRNA changes that drive cell differentiation in the normal colonic crypt [19]. SC, stem cells; Abs., absorptive; Sec., secretory/deep crypt secretory cells/goblet; EEC, enteroendocrine; Ent, enterocytes, TuftC, tuft cells. The pathway analysis of the 134 genes shows a significant enrichment in TGF-beta signaling pathway ($P_{adj}=0.032$), BioPlanet 2019. **C** Left panel, heatmap displaying pathway analysis of coding mutations, exon skipping signature (All, Alternative, Constitutive, DD134, ES96), and of the exon skipping + coding mutations; colors represent enrich pathway P_{adj} , $-\log_{10}$. Enrichment was realized with EnrichR BioPlanet 2019. Terms were sorted by P_{adj} (< 0.05) and regroup first in pathways (rows, right) then in groups of pathways (rows, left). Columns were sorted by enriched condition parameters. A more detail heatmap showing the different terms is available in Fig. S12. Right panel, interactions between genes related to coding repeat mutations and/or genes with alterations in the splicing of genes involved in Hippo signaling pathways which play a role in CRC and in cell differentiation in the colonic crypt (adapted from KEGG: hsa04390). **D** Left panel, number of neoepitopes potentially represented by HLA-MHC class I or II for each patient ($n=101$ CRC MSI). Middle panel, boxplot displaying the number of unique neoepitopes in average per transcript and per patient. Right panel, boxplot displaying the number of neoepitopes showing high affinity for HLA-MHC class I or II for each patient. **E** Constitutive exon skipping with a frameshift (ES96) exposing a neoantigen tail; genes were ordered by mutational frequency in the early stage. **F** Overlapping signature of 125 exons (120 genes) containing common exons between the U2AF-S34F signature and the MSI splicing signature

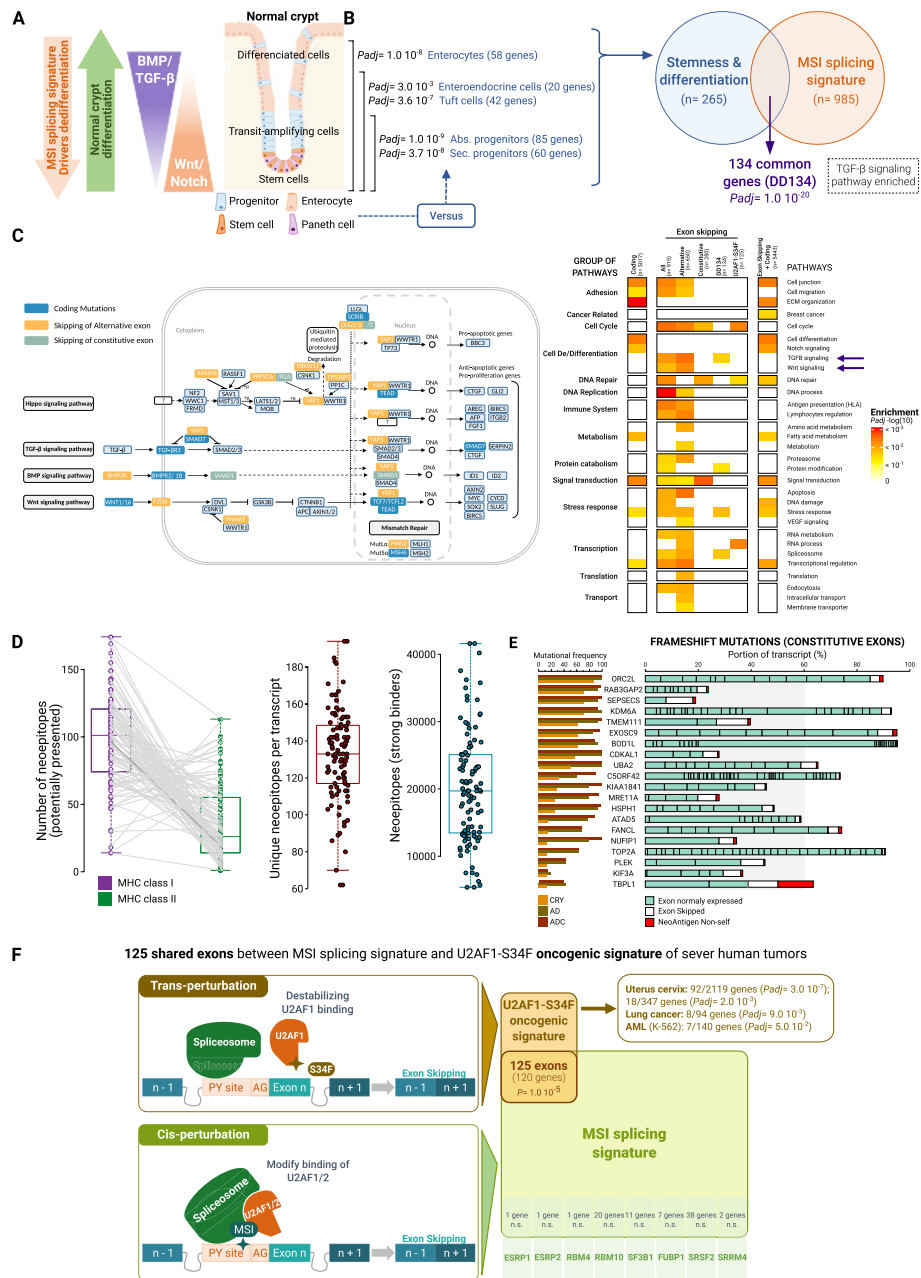


Fig. 5 (See legend on previous page.)

The MSI splicing signature also generates early immunogenic aberrant frameshifts at constitutive exons

In addition, we investigated whether splicing defects affecting constitutive exons in mRNAs might have an early immunogenic impact since this kind of exon skipping is not physiologically observed in normal colonic tissue. Considering the HLA-MHC status of each patient (CRC only) and the presentation of aberrant peptides generated in this setting, many immunogenic neopeptides were predicted to be presented by the patient’s tumor cells (Fig. 5D). These data strongly suggest an immunogenic nature of the MSI splicing signature, and we further examined its effect on the early stages of MSI CRC

tumorigenesis. A significant number of these constitutive immunogenic exons were present in dMMR crypts (20/27, 74.1%), and 10 of them had a high mutational frequency ($MF \geq 50\%$) (Fig. 5E). These results suggest that immunogenicity in the early stage can stimulate immune infiltration in the normal colonic mucosa, as recently reported [37].

The aberrant splicing events observed in MMR-deficient CRC are similar to those induced by oncogenic dysfunction of the spliceosomal factor U2AF1, which promotes blockade of cell differentiation in several human cancer types

We next investigated whether exon skipping events in MSI CRC could be shared with oncogenic signatures observed following mutations of core spliceosomal proteins or associated *trans*-acting RNA splicing factors in other cancer types (Fig. 5F left panel). The signature ($n=985$ exons) was compared to results from 20 publications that described the exon signature of 9 RNA splicing factors (ESRP1, ESRP2, FUBP1, RBM10, RBM4, SF3B1, SRRM4, SRSF2, U2AF1) in different cancer types. Strikingly, the MSI splicing signature was significantly enriched in exons that were skipped following U2AF1 dysfunction in several cancer types [38–40] (Fig. 5F right panel, Additional file 1: Fig. S14 and Additional file 7: Table S6). This enrichment corresponded to a total of 125 unique exons and the results concur with the known role of mutations in U2AF1 in cancer, particularly U2AF1 S34F in lung carcinoma, uterus cervix carcinoma, and acute myeloid leukemia [41, 42]. The present results suggest that, mechanistically, the *cis*-acting, MSI-driven splicing pathway promotes tumor development by mimicking the oncogenic impact of U2AF1 dysfunction on several gene targets (Fig. 5F left panel). This underlines further that aberrant splicing due to MSI can promote tumor development through *cis*-somatic MS deletions that may mimic U2AF1 perturbation of function and result in skipping of numerous exons, thereby defining a U2AF1-like MSI-driven oncogenic signature (see also Additional file 1: Fig. S15 for the influence on exon skipping of length and deletion size of these intronic MS). In contrast, the impact of additional mutations that occurred outside MS at the donor (GT) or acceptor (AG) sites on exon exclusion levels was shown to be very low (Additional file 1: Fig. S16).

Discussion

MSI is a common type of genetic instability in human solid tumors, and approximately 700,000 new cases of MSI cancers are diagnosed worldwide every year. To date, this mode of instability and its pathophysiological consequences regarding tumorigenesis have been mostly reported at the DNA coding level [3, 10, 16, 43–46]. Here, we report for the first time the early impact of MSI on gene splicing in both the untransformed and transformed dMMR colonic mucosa through a newly identified MSI-driven non-coding genomic pathway. In brief, coding MS mutations are frameshift and are usually observed as 1 or 2 bp deletions in short DNA repeats and at low/moderate frequencies [8]. These mutations invariably lead to truncation of the corresponding protein in tumor cells, resulting in aberrant proteins which may have lost their function. Although a role in tumorigenesis has been established for several coding MS mutations, they mostly result in heavy tumor infiltration with cytotoxic T cells and hence an improved prognosis due to their neo-antigenic properties [11, 47]. Our data suggest that these coding MSI variants occur late in tumor development and are thus rarely observed during the

early steps of MSI tumorigenesis. In contrast, the MS often found in intronic sequences around the PY are shown here to undergo frequent deletions of one or several bp in MSI cancer cells but also in yet untransformed dMMR colonic crypts, particularly the long MS contained in the *cis*-regulatory intronic polypyrimidine tract. Strikingly, these MSI-driven noncoding changes early deregulates at high frequencies the expression of dozens of exons which combine with many other exon skipping events unlikely to be directly linked to MSI at the PY site to give rise to a broad splicing signature characteristic of MSI CRC, involving hundreds of exons in all. Other MSI-driven *cis*-scenarios investigating mutations of microsatellites endowed with functional activity in splicing outside the PY, like in exonic (ESE, ESI) and intronic (SSE, SSI) splicing enhancers and silencers, will have to be developed in future studies to illustrate why MSI-driven PY mutations explain only about 10% of the exon skipping events observed in these cancers [48].

Overall, our data highlight that the number of exon skipping events in established MSI colon tumor, which are in part MSI-driven at the PY, is approximatively similar to the number of somatic mutations in coding MS. Importantly, alternative exons are shown here to constitute the main early target of this process in MSI CRC, probably due to their greater susceptibility to undergo skipping at the intron/exon junction. Though the nature of “sequencing used in our single-cell RNA study limits our capacity to specifically classify alternative splicing (AS) events, such as exon skipping, we confirm this close link between the MSI phenotype of tumor clones and incrementation of splicing events in an independent, single-cell level dataset. Thus, the splicing pathway we highlight in this study mainly deregulates the fine tuning of many self-protein isoforms that are normally tightly regulated under physiological conditions and govern the process of cell fate determination in the normal colonic mucosa [17–19]. Therefore, though our data only report indirect evidence for such a scenario, they highly suggest that this would drastically impair the differentiation of dMMR colonic cells by notably targeting TGFβ signaling. Besides, a number of constitutive exons are however also targeted by this MSI pathway in MSI cancer cells which may also initiate immunization mechanisms prior to cell transformation. These data are especially meaningful in the context of recent studies where elevated mucosal T cell infiltration was demonstrated in Lynch patients even in the absence of cancer [37].

Of pathophysiological interest also is that our aberrant exon skipping signature is best characterized by its similarity to that of the oncogenic U2AF1S34F mutation in several other non-MSI cancer types [38–40]. Without resolving this issue that will have to be the subject of future in-depth, more specialized studies, we bring here functional data which examine the molecular basis of a subset of exon skipping belonging to this signature, i.e., those that are MSI-driven and associated with PY mutations in MSI colorectal cancer cells. Regardless of these mechanistic aspects, the similarity of our RNA splicing signature with the one associated with U2AF1S34F mutation constitutes an important argument for the oncogenic and transforming role of exon skipping events in the dMMR colonic crypt, well before the burst of coding mutations that arises much later in transformed MSI cancer cells. The S34F mutation in U2AF1 is commonly found in myelodysplastic syndromes and secondary acute myeloid leukemia [49]. In line with our findings, it is responsible for blocking erythroid differentiation and causing aberrant alternative splicing in hematopoietic progenitors associated with different disease phenotypes [49, 50].

A number of targets common to both U2AF1 and MSI splicing signatures (e.g., ADAM10 [51], CENP-E [52]) may underpin this functional impact. Strikingly, MSI CRC are frequently characterized by dedifferentiated histopathology [53]. From all our results, we speculate that the RNA splicing defects we report in this work induce early dedifferentiation of multiple dMMR clones in the crypt and then the transformation of some of them at the apex of cancer initiation. It thus appears, that pro-differentiation or splice-correcting agents might be considered as drugs of choice for the early treatment of MSI lesions and more generally for MMR-deficient cancers. These drugs could be proposed to improve the personalized therapy of MSI cancers, at a time when precision medicine for MSI cancers using immune checkpoint inhibitors is growing fast, with the need however to find new targets in combination to overcome resistance to these new drugs, as recently illustrated by our team and others [16].

There are some limitations to this study that have been already highlighted. Mechanistically, though a number of somatic mutations at the PY are likely to participate to the initiation of colon cancer, we have not precisely unraveled in this work their exact consequences on the binding of spliceosome factors to RNA. Besides, alternative scenarios have to be examined in future studies since MSI at the PY only explain a minority of the exon skipping we report here in MSI CRC, with yet a poor understanding of the link between these splicing anomalies and those due to the U2AF1 mutation. From a pathophysiological point of view, some might object that the number of splicing anomalies we report here in MSI CRC is not exceptionally high. This is partly due to the fact that we have willingly chosen to detect only the events with a frequency exceeding 5% in the tumor bulk. Note worthily, this process is here validated in several tumor series including tumor samples from colorectal and other epithelial tissue locations, suggesting that it probably plays a universal major role in MSI tumorigenesis. The next step will be to more accurately describe these splicing anomalies in pan-cancer, using in particular long-read sequencing technology to establish their combinations precisely in tumors, and earlier as we initiated here in pre-tumor lesions.

Conclusions

In this work, we show that alternative splicing is altered partly due to MSI in the intestinal crypt, and that this process is crucial for cell dedifferentiation and initiation of MSI tumorigenesis far before the MSI-driven coding pathway.

Methods

Samples and patients

This study included an experimental cohort composed of 101 MSI patients and 32 MSS CRC patients for whom both tumor tissue and matched colonic mucosa were available. The data were obtained from a retrospectively enrolled cohort of patients who underwent surgical resection of mainly stage II–III CRC from 2004 to 2015 at Saint Antoine Hospital, Paris, France (Cohort Microsplicother, NI14027HLJ). Additionally, 22 pre-cancerous tissues—adenoma ($n = 14$ samples) and dMMR crypts ($n = 8$ samples, a pool of 10 crypts foci for each patient)—associated with their normal mucosa ($n = 22$), i.e., MMR-proficient crypts foci. The data were obtained from a retrospectively cohort of patients who underwent surgical resection of mainly stage II–III CRC from 2013 to 2020

at Lille University Hospital, France. We included patients who presented with a tumor showing loss of MLH1 or MSH2 expression at pathological examination. MSI/dMMR status was identified prospectively at diagnosis using the pentaplex PCR method and immunohistochemistry as previously described [7, 35]. Patients who received preoperative chemotherapy and/or radiation therapy were excluded. Extensive clinical data were available for all MSI CRC patients. This study was approved by our institutional review board/ethics committee (Sorbonne University), and informed consent was obtained. This research was performed in accordance with the Declaration of Helsinki.

RNA sequencing

Before RNA extraction, frozen tissue sections were lysed in QIAzol Lysis Reagent using a Tissue Lyzer (Qiagen). After chloroform separation, RNA extraction was performed using the miRNeasy Mini Kit (Qiagen) with a QIAcube instrument (Qiagen). The manufacturer's instructions were followed, and DNase treatment was included in the protocol. RNA concentrations were evaluated with a Nanodrop 2000C (Thermo Fischer Scientific) spectrophotometer. RNA integrity was assessed on a Bioanalyzer 2100 (Agilent) using the RNA 6000 Nano Kit. The average RINs (RNA Integrity Number) calculated for tumor and normal adjacent tissues were equal 8.1 and 7.2, respectively. Downstream RNA sequencing experiments were performed on selected pairs of samples, each with $RIN \geq 7$ and a DNA yield $\geq 2.5 \mu\text{g}$.

Stranded mRNA sequencing was performed by the "Centre National de Recherche en Génomique Humaine, Institut de Biologie François Jacob, CEA." After complete RNA quality control of each sample (quantification in duplicate and RNA6000 Nano Lab-Chip analysis on Bioanalyzer from Agilent), libraries were prepared using the "TruSeq Stranded mRNA Library Prep Ki" from Illumina according to the manufacturer's instructions and with an input of 1 μg (selection of poly(A) RNAs). Library quality was checked by Bioanalyzer analysis, and sample libraries were then pooled before sequencing to reach the expected sequencing depth. Sequencing was performed on an Illumina HiSeq 2000 sequencer as paired-end 100 bp reads using Illumina sequencing reagents, 50 M reads. Libraries were pooled with 4 samples per lane. FASTQC v.0.11.7 was used for sequencing quality control, and the reads were aligned against the reference genome hg37 (GRCh37) using TopHat2.

Identification of exon skipping events

MSIsensor [54] (scan method) 0.5 was used to identify genomic MS and annotation was performed by ANNOVAR [55]. We identified all exons flanked by an MS near an intron/exon junction (up to 50 bp into the intronic sequence) and designed a custom junction FASTA reference (50 bases on each side of the junction). The identification of reads spanning an exon-exon junction was performed by BLASTN alignment against our custom reference for each 3'-exon flanked by an MS (3' Exon-FMS). A spanning read was considered if its sequence matched at least 8 bases on one side of the junction and the remainder of the sequence strictly matched the other side of the junction. Percent spliced in (PSI) [28] scores ranging from zero to one were calculated by $(\text{inclusion reads}) / (\text{inclusion} + \text{exclusion reads})$ and considered when the event was covered by at least 5 reads (inclusion or exclusion reads).

For each 3' Exon-FMS, the PSI of MSI or MSS tumors was compared to that of normal tissues. Adjusted *P*-values were obtained using Benjamini–Hochberg correction of two-sided *t*-test ($p\text{-adjust} < 0.05$). In parallel, we also compared the PSI of MSI tumors to the PSI of MSS tumors ($p\text{-adjust} < 0.05$). For 3' Exon-FMS identified by the two-sided *t*-test adjusted as skipped, we compared the PSI of each tumor or PSI of each normal matched tissue. The exons considered as skipped had a PSI fold change greater than 1.5 between tumor and normal. Finally, only exons with a mutation frequency >5% in this cohort were considered. Integrative Genomics Viewer (IGV) was used to generate Sashimi plot. From the 133 available healthy tissues, we considered an exon as an alternative exon if at least 5% of the tissues presented 3 or more reads of exon exclusion. We did not replicate the same analysis on publicly available cohorts because we recently provided clear evidence that the performance of MSISensor for the detection of MSI vs. MSS CRC has to be improved. This was shown in large cohorts of mCRC and nmCRC samples that were previously confirmed as MSI/dMMR or MSS/pMMR by IHC and MSI-PCR methods performed in large, specialized test centers [27].

Tissue recovery and DNA/RNA extraction

Frozen tissues (adenocarcinoma) and formalin-fixed and paraffin-embedded tissues (adenomas and crypts) were recovery by punch technics (adenocarcinomas and adenomas) or by an innovative method developed by our team and based on laser microdissection (LMD7000 Laser, Leica) assisted by IHC (hMSH2 1/125 FE11 clone, Calbiochem) (MMR-deficient and proficient crypts). DNA was purified with the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen) for frozen tissue and AllPrep DNA/RNA FFPE Kit (Ozyme) for FFPE cores, according to the manufacturer's instructions. Their concentrations were determined by a NanoDrop ND-1000 spectrophotometer at 260/280 nm (NanoDrop Technologies, Inc.) and by the Qubit Fluorometer (Qubit® dsDNA HS Assay Kit). The quality and integrity of the nucleic acids were analyzed with a TapeStation 2200 system (Agilent).

Exome data analyses

For adenocarcinomas, the whole exome sequencing procedure was performed following the manufacturer's recommendations (SureSelect Human Exon Kit v5, 75 MB; Agilent). The generated reads were mapped to reference genome hg19 (GRCh37).

For adenomas and crypts, the whole exome sequencing procedure was performed following the manufacturer's recommendations (KAPA HyperExome 45 Mb, Roche, Illumina NovaSeq 6000 S1 Reagent Kit – 2 × 100 cycles), and we used the Unique Molecular Identifier (UMI), following the manufacturer's recommendations. The exome data analysis was first performed using fastp with a minimum read length of 15 and mapped to the reference genome hg19 (GRCh37) using bwa v0.7.17. Overall, variant calling (single-nucleotide variants and insertions/deletions) was performed using MuTect-2 tool from GATK v4.1.6 [56]. The somatic mutations were filtered as follows: total reads > 10, somatic score > 3, mutated allele frequency in tumor tissue ≥ 5%, mutated allele count in tumor tissue ≥ 3, mutated allele frequency in normal tissue < 4%. A somatic mutation observed at the donor (GU) or acceptor (AG) site is considered to affect splicing if it induces a delta of PSI > 0.1. To analyze mutations extensively at microsatellite sequence

sites, mononucleotide MS with lengths greater than or equal to 5 bp were considered if they were covered with at least 10 mapping reads in both normal and tumor paired samples. MSIsensor 0.5 was initially used to generate the read count distribution for each MS. The total number of reads covering each MS was normalized (arbitrary value of 100) in tumor and healthy matched tissue. For each MS, the normalized value of the reads in healthy tissue was subtracted from that in the tumor tissue. If the ratio difference between normal and tumor tissues was > 10% for deletion the MS was identified as mutated. The deletion size considered for each MS corresponds to the largest deletion value with a ratio of difference between normal and tumor > 10%. The branch sites were predicted using SVM-BPfinder with the default parameters [57].

Intronic microsatellite mutations analyses were allowed (from WES) because of the sufficiently high sequencing depth and good median coverage of the flanking sequences of the exons (Additional file 1: Fig. S17 and Additional file 8: Table S7).

DNA preparation and mutation analyses (functional)

Genomic DNA was extracted using a standard phenol–chloroform procedure and was assessed for integrity and quantity with a Bioanalyzer (Agilent, Les Ulis, France). Primers specific for each locus were used to amplify the repeat and short flanking sequences from template DNA using PCR (cycling conditions, parameters and primers sequences available on request). Amplified PCR products were run on an Applied Biosystems PRISM 3100 Genetic Analyzer automated capillary electrophoresis DNA sequencer. Allelic sizes were estimated using gene mapper software (Applied Biosystems).

Cell culture and transfection

We have deliberately used MSI and MSS gastric cancer cellular models to better illustrate the broad scope of our results, which have the potential to apply to other primary MSI cancers than CRC (see also Fig. 3E our results in primary MSI GC and EC). These cellular models have been selected without preconceived ideas, and we have not excluded any others. This series corresponds to cellular models commonly investigated in the literature, which we have historically already used extensively in our laboratory. Cell lines were grown at 37 °C under a humidified atmosphere of 5% CO₂ in DMEM (Dulbecco's modified Eagle's medium) (Lonza, Saint Beauzire, France), containing 4.5 g L⁻¹ glucose and supplemented with 5% fetal bovine serum without antibiotics (Gibco-Invitrogen, Cergy-Pontoise, France). Cells were plated at a density of 75.10⁴ cells per well in 6-well plates 24 h before transfection. Minigenes (0.75 µg per well) were transfected with Lipofectamine 2000 or 3000 reagent according to the manufacturer's recommendation (Invitrogen). After 24 h, the cells were harvested, washed in PBS, and used to prepare total RNA. Cell lines were free of mycoplasma contamination.

Minigene constructions

PMSG-1-PolR2G hybrid minigenes have been described previously [34]. The genomic segment encompassing the sequences of interest, containing roughly 150 bp of the upstream intron, the next exon, and 150 bp of the downstream intron, was amplified by PCR from tumor or non-tumor genomic DNA. The resulting plasmids harboring wild

type or mutant sequences were verified by sequencing. Plasmid batches for transfection experiments were prepared according to the manufacturer's procedure (Qiagen).

RNA isolation and RT-PCR analyses

Total RNA was extracted with Tri-Reagent (Ambion, Les Ulis, France) or NucleoSpin RNA kit (Macherey–Nagel), according to the manufacturer's instructions. RNA was quantified by spectrophotometry, and the integrity was analyzed by gel electrophoresis. One microgram of RNA was reverse-transcribed with M-MLV reverse transcriptase (Biolabs) and random primers or the High-Capacity cDNA Archive Kit (Applied Biosystems). The resulting cDNA was amplified.

RNA immunoprecipitation (RIP)

RIP assays were performed according to the manufacturer's instructions (Millipore, EZ-Nuclear RIP Kit (Catalog No. 17–10,521)). Briefly, cells were grown in 150-mm Petri dishes. Magnetic beads and anti-U2AF2 or IgG control antibodies were added to the nuclear protein extracts and put on a rotating wheel at 4 °C overnight to immunoprecipitate protein-RNA complexes. Beads were washed several times and eluted protein-RNA complexes were dissociated (with 1% SDS and proteinase K) at 60 °C for 30 min with gentle shaking. RNA was then purified using TRIzol, precipitated, resuspended in water, and then treated with DNase I. The DNase was inactivated, and the RNA extracted and used for RT-PCR experiments (High-Capacity cDNA Archive Kit and PCR-amplified with Go Taq polymerase (Promega)).

Gel shift assay analyses

Two gel retardation assays were performed using 5' end-labeled RNA wild type and mutant probes. Samples (4 µg) of nuclear protein extracts were incubated with 2 nmoles of the labeled probe at room temperature for 25 min and run for 1 h in non-denaturing 1.5% agarose gels. The gels were analyzed by fluorescence illumination on an Odyssey™ scanner device (LI-COR). The Cy5.5 signal appeared in red, while the HF750 signal appeared in green.

Immunoblot analyses

Cells were washed in PBS and lysed on ice in cold lysis buffer (RIPA buffer, Thermo Scientific, France) containing protease and phosphatase inhibitors (Halt Protease and Phosphatase Inhibitor Cocktail, Pierce). Proteins were separated by PAGE and transferred following standard protocols before analysis with an Odyssey Infrared Imaging System. Primary antibodies against the following molecules were used: Hsp110 (ab24503 dilution 1:100; Abcam, Cambridge, UK) and actin (926–42,210 dilution 1:5000; Licor, The Netherlands). IRDye680 and IRDye800 secondary antibodies were used (926–68,021 and 926–32,211, respectively, dilution 1:15,000, Licor).

Functional analyses

Genes were analyzed using Enrichr to determine the signaling pathways (BioPlanet 2019) that were enriched in the mutated genes [58].

MHC class I and class II neoepitope prediction

Neoepitopes were predicted through the Ideation@SiRIC pipeline, which integrates various software packages. First, seq2HLA was employed to determine MHC class I and class II types [59], using default parameters for 101 normal-WES preprocessed fastq files via fastp. Next, the alternative transcripts were processed with the pVACbind function from the pVACtools toolkit for neoantigen prediction [60]. In each pVACbind run, NetMHC [61], NetMHCpan, and NetMHCIIpan [62] algorithms, encapsulated in the pVACtools, were employed for epitope prediction. The epitope length was set to 8–10 amino acids for class I and 15 for class II presentation, with default parameters for all other settings. Predicted neoepitopes were filtered based on median affinity binding ≤ 500 nM, representing strong binders. In instances where multiple strong binders were predicted for the same alternative transcript, a unique neoepitope per transcript was defined as the transcript with the lowest binding affinity score. The count of unique neoepitopes per transcript was determined as the overall number of unique filtered transcript peptide sequences per patient.

Single cell analysis

Single cell RNA-seq (scRNA-seq) data were obtained from the Pelka et al. study [26] via dbGaP (phs002407.v1.p1). Raw FASTQ files were aligned to the GRCh38 reference genome using Cell Ranger v2.1.1, to generate BAM files. These BAM files were used as input for STARsolo (bioRxiv, 2021.2005.2005.442755), component of STAR v2.7.8a, to generate gene expression count matrices. The SingCellaR package [63] was then used to filter for high quality cells based on UMI counts and the number of detected genes. The preprocessed data were further analyzed using Scanpy [64] for dimensionality reduction with uniform manifold approximation and projection (UMAP) [65] and plotting.

Cell types were identified using Census (<https://doi.org/10.1101/2022.10.19.512926>) [29], which implements a collection of hierarchically organized gradient-boosted decision tree models for cell type classification. Specifically, we used Census to identify cancer cells. Additionally, we used CellTypist [30] and its pretrained “*Cells_Intestinal_Tract.pkl*” model for automated cell type prediction based on gene expression profiles.

Splice junction counts were generated from the FASTQ files using STARsolo. The junction counts and reference GTF files were input into the MARVEL package [31] to create an R object for downstream analyses. MARVEL was used to annotate junctions, filter unannotated and multimapped junctions, and estimate percent spliced-in (PSI) values directly from junction reads. Finally, we performed splicing analysis to compare PSI values between pseudobulk samples using the permutation-based approach of MARVEL [66, 67] to compare transit amplifying (TA) cells based on the MSI/MSS phenotype.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03340-5>.

Additional file 1: Supplementary Figs. S1 to S17.

Additional file 2: Table S1.

Additional file 3: Table S2.

Additional file 4: Table S3.

Additional file 5. Table S4.
 Additional file 6. Table S5.
 Additional file 7. Table S6.
 Additional file 8. Table S7.
 Additional file 9. Uncropped Western Blot.
 Additional file 10. Uncropped Western Blot.
 Additional file 11. Uncropped Western Blot.
 Additional file 12. Review history.

Acknowledgements

We thank the CEPH-Biobank and CNRGH teams for their technical expertise. We thank the Tumorothèque CRB Cancer, Hôpital Tenon, APHP Sorbonne université, Paris, France. We would like to thank Christophe Rachez for his technical help. We thank Ms Prigent who is the technical referent of Histomics (ICM, Sorbonne University).

Peer review information

Ulf Schmitz and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 12.

Authors' contributions

All the authors have made substantial contributions (i) to the conception or design of the work; (ii) the acquisition, analysis, or interpretation of the data; and (iii) the drafting or substantial revision of the work. All authors have approved the submitted version. They have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even those in which the author was not personally involved, are appropriately investigated and resolved and that the resolution is documented in the literature.

Funding

This work was supported by grants from SIRIC CURAMUS and the "Ligue Nationale Contre le Cancer." A. Duval's team is labeled by the French National League against Cancer. The sponsor was Assistance Publique – Hôpitaux de Paris (Département de la Recherche Clinique et du Développement). The study was funded by a grant from Programme Hospitalier de Recherche Translationnelle en Cancérologie – PHRT-K 14056 (Ministère de la Santé). This work was also supported by the France Génomique National infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09). All contributors from Brest U1078 INSERM unit were supported by grants from the Ligue contre le cancer (departments 29 and 44) and by grants from the GRAMMY EraPerMed European research program.

Availability of data and materials

Exome sequencing and RNA sequencing data have been deposited in the European genome–phenome archive. The data deposit is available on EGA with the accession number EGAS00001004863 [68]. All the scripts for analyzing and pre-processing raw data sequences are available under GPL-3.0 license at GitHub: https://github.com/CRSA-MSI/pipelines_and_scripts_Jonchre_Montemont_et_al_2024 [69] and in Zenodo: <https://doi.org/10.5281/zenodo.12706170> [70].

Declarations

Ethics approval and consent to participate

This study was approved by our institutional review board/ethic committee (Sorbonne University), and informed consent was recorded. Research has been performed in accordance with the Declaration of Helsinki.

Competing interests

PDLG and AJ (authors) declare a conflict of interest with the Genosplice company.

Author details

¹Sorbonne Université, INSERM, Unité Mixte de Recherche Scientifique 938 and SIRIC CURAMUS, Centre de Recherche Saint-Antoine, Equipe Instabilité Des Microsatellites Et Cancer, Equipe Labellisée Par La Ligue Nationale Contre Le Cancer, 75012 Paris, France. ²INSERM, UMR 1078, Université de Brest, Génétique Génomique Fonctionnelle Et Biotechnologies, Etablissement Français du Sang, F-29200 Brest, France. ³CHU de Brest, Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest F-29200, France. ⁴Laboratory for Genomics, Fondation Jean Dausset–CEPH (Centre d'Etude du Polymorphisme Humain), Paris, France. ⁵Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057 Evry, France. ⁶Genosplice, Paris, France. ⁷Programme "Cartes d'Identité Des Tumeurs, Ligue Nationale Contre Le Cancer, Paris, France. ⁸Department of Pathology, Sorbonne Université, AP-HP Sorbonne Université Hôpital Saint-Antoine, 47-83 Boulevard de L'hôpital, 75012 Paris, France. ⁹Department of Digestive Surgery, Sorbonne Université, AP-HP, Hôpital Saint-Antoine, Paris, France. ¹⁰Sorbonne Université, Inserm, CNRS, UMR S 1127 and SIRIC CURAMUS, Institut du Cerveau Et de La Moelle Épineière, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neurologie 2 Mazarin, Paris, France. ¹¹Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau Et de La Moelle Épineière, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neuropathologie Laboratoire Escourrolle,

Paris, France. ¹²Department of Medical Oncology, Sorbonne Université, AP-HP, Hôpital Saint-Antoine, Paris, France.
¹³Genetics Department, AP-HP, Sorbonne Université, Paris, France.

Received: 24 October 2023 Accepted: 22 July 2024

Published online: 06 August 2024

References

- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–74.
- Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW, Vogelstein B, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*. 1995;268:1336–8.
- Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, Reed JC, Perucho M. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science*. 1997;275:967–9.
- Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Powell SM, Jen J, Hamilton SR, et al. Clues to the pathogenesis of familial colorectal cancer. *Science*. 1993;260:812–6.
- Branch P, Aquilina G, Bignami M, Karan P. Defective mismatch binding and a mutator phenotype in cells tolerant to DNA damage. *Nature*. 1993;362:652–4.
- Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*. 1993;75:1027–38.
- Svrcek M, Lascols O, Cohen R, Collura A, Jonchere V, Flejou JF, Buhard O, Duval A. MSI/MMR-deficient tumor diagnosis: which standard for screening and for diagnosis? Diagnostic modalities for the colon and other sites: differences between tumors. *Bull Cancer*. 2019;106:119–28.
- Jonchere V, Marisa L, Greene M, Virouleau A, Buhard O, Bertrand R, Svrcek M, Cervera P, Goloudina A, Guilleum E, et al. Identification of positively and negatively selected driver gene mutations associated with colorectal cancer with microsatellite instability. *Cell Mol Gastroenterol Hepatol*. 2018;6:277–300.
- Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*. 2013;155:858–68.
- Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. *Cancer Res*. 2002;62:2447–54.
- Lothe RA, Peltomaki P, Meling GI, Aaltonen LA, Nystrom-Lahti M, Pylkkanen L, Heimdal K, Andersen TI, Moller P, Rognum TO, et al. Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. *Cancer Res*. 1993;53:5849–52.
- Marisa L, Svrcek M, Collura A, Becht E, Cervera P, Wanherdrick K, Buhard O, Goloudina A, Jonchere V, Selves J, et al. The balance between cytotoxic T-cell lymphocytes and immune checkpoint expression in the prognosis of colon tumors. *J Natl Cancer Inst*. 2018;110:68–77.
- Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med*. 2015;372:2509–20.
- Cohen R, Bennouna J, Meurisse A, Tournigand C, De La Fouchardiere C, Tougeron D, Borg C, Mazard T, Chibaudel B, Garcia-Larnicol ML, et al. RECIST and iRECIST criteria for the evaluation of nivolumab plus ipilimumab in patients with microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer: the GERCOR NIPICOL phase II study. *J Immunother Cancer*. 2020;8:e001499.
- Cohen R, Svrcek M, Dreyer C, Cervera P, Duval A, Pocard M, Flejou JF, de Gramont A, Andre T. New therapeutic opportunities based on DNA mismatch repair and BRAF status in metastatic colorectal cancer. *Curr Oncol Rep*. 2016;18:18.
- Ratovomanana T, Nicolle R, Cohen R, Diehl A, Siret A, Letourneur Q, Buhard O, Perrier A, Guilleum E, Coulet F, et al. Prediction of response to immune checkpoint blockade in patients with metastatic colorectal cancer with microsatellite instability. *Ann Oncol*. 2023;34(8):703–13.
- Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*. 2017;18:437–51.
- Fiszbein A, Kornbliht AR. Alternative splicing switches: important players in cell differentiation. *Bioessays*. 2017;39:1600157.
- Habowski AN, Flesher JL, Bates JM, Tsai CF, Martin K, Zhao R, Ganesan AK, Edwards RA, Shi T, Wiley HS, et al. Transcriptional and proteomic signatures of stemness and differentiation in the colon crypt. *Commun Biol*. 2020;3:453.
- Buhard O, Lagrange A, Guilloux A, Colas C, Chouchene M, Wanherdrick K, Coulet F, Guilleum E, Dorard C, Marisa L, et al. HSP110T17 simplifies and improves the microsatellite instability testing in patients with colorectal cancer. *J Med Genet*. 2016;53:377–84.
- Qu H, Wang Z, Zhang Y, Zhao B, Jing S, Zhang J, Ye C, Xue Y, Yang L. Long-read nanopore sequencing identifies mismatch repair-deficient related genes with alternative splicing in colorectal cancer. *Dis Markers*. 2022;2022:4433270.
- Dorard C, de Thonel A, Collura A, Marisa L, Svrcek M, Lagrange A, Jego G, Wanherdrick K, Joly AL, Buhard O, et al. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. *Nat Med*. 2011;17:1283–9.
- Giannini G, Rinaldi C, Ristori E, Ambrosini MI, Cerignoli F, Viel A, Bidoli E, Berni S, D'Amati G, Scambia G, et al. Mutations of an intronic repeat induce impaired MRE11 expression in primary human cancer with microsatellite instability. *Oncogene*. 2004;23:2640–7.
- Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44:W90–97.
- Romanelli MG, Diani E, Lievens PM. New insights into functional roles of the polypyrimidine tract-binding protein. *Int J Mol Sci*. 2013;14:22906–32.

26. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, Bejnood A, Dionne D, Ge WH, Xu KH, et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*. 2021;184(4734–4752): e4720.
27. Ratovomanana T, Cohen R, Svrcek M, Renaud F, Cervera P, Siret A, Letourneur Q, Buhard O, Bourgoin P, Guillem E, et al. Performance of next-generation sequencing for the detection of microsatellite instability in colorectal cancer with deficient DNA mismatch repair. *Gastroenterology*. 2021;161(814–826):e817.
28. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr Protoc Hum Genet*. 2015;87:11.16.1–11.16.14.
29. Bassel Ghaddar B, De S. Census: accurate, automated, deep, fast, and hierarchical scRNA-seq cell-type annotation. *bioRxiv* 2022.10.19.512926; <https://doi.org/10.1101/2022.10.19.512926>.
30. Dominguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, Howlett SK, Suchanek O, Polanski K, King HW, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*. 2022;376:eabl197.
31. Wen WX, Mead AJ, Thongjuea S. MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data. *Nucleic Acids Res*. 2023;51:e29.
32. Yoshida H, Park SY, Sakashita G, Nariai Y, Kuwasako K, Muto Y, Urano T, Obayashi E. Elucidation of the aberrant 3' splice site selection by cancer-associated mutations on the U2AF1. *Nat Commun*. 2020;11:4744.
33. Bokhari A, Jonchere V, Lagrange A, Bertrand R, Svrcek M, Marisa L, Buhard O, Greene M, Demidova A, Jia J, et al. Targeting nonsense-mediated mRNA decay in colorectal cancers with microsatellite instability. *Oncogenesis*. 2018;7:70.
34. Fichou Y, Gehannin P, Corre M, Le Guern A, Le Marechal C, Le Gac G, Ferec C. Extensive functional analyses of RHD splice site variants: Insights into the potential role of splicing in the physiology of Rh. *Transfusion*. 2015;55:1432–43.
35. Cohen R, Hain E, Buhard O, Guilloux A, Bardier A, Kaci R, Bertheau P, Renaud F, Bibeau F, Flejou JF, et al. Association of primary resistance to immune checkpoint inhibitors in metastatic colorectal cancer with misdiagnosis of microsatellite instability or mismatch repair deficiency status. *JAMA Oncol*. 2019;5:551–5.
36. Beumer J, Clevers H. Cell fate specification and differentiation in the adult mammalian intestine. *Nat Rev Mol Cell Biol*. 2021;22:39–53.
37. Bohaumilitzky L, Kluck K, Huneburg R, Gallon R, Nattermann J, Kirchner M, Kristiansen G, Hommerding O, Pfuederer PL, Wagner L, et al. The different immune profiles of normal colonic mucosa in cancer-free Lynch syndrome carriers and Lynch syndrome colorectal cancer patients. *Gastroenterology*. 2022;162(907–919):e910.
38. Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, Zhou J, Qiu J, Jiang L, Li H, et al. Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol*. 2014;21:997–1005.
39. Brooks AN, Choi PS, de Waal L, Sharifnia T, Imielinski M, Saksena G, Peadarallu CS, Sivachenko A, Rosenberg M, Chmielicki J, et al. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS ONE*. 2014;9:e87361.
40. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, Bradley RK. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res*. 2015;25:14–26.
41. Uhlen M, Zhang C, Lee S, Sjostedt E, Fagerberg L, Bidkhori G, Benfais R, Arif M, Liu Z, Edfors F, et al. A pathology atlas of the human cancer transcriptome. *Science*. 2017;357. <https://doi.org/10.1126/science.aan2507>.
42. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347:1260419.
43. Duval A, Gayet J, Zhou XP, Iacopetta B, Thomas G, Hamelin R. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. *Cancer Res*. 1999;59:4213–5.
44. Duval A, Reperant M, Compoint A, Seruca R, Ranzani GN, Iacopetta B, Hamelin R. Target gene mutation profile differs between gastrointestinal and endometrial tumors with mismatch repair deficiency. *Cancer Res*. 2002;62:1609–12.
45. Hamelin R, Chalastanis A, Colas C, El Bchiri J, Mercier D, Schreurs AS, Simon V, Svrcek M, Zaanen A, Borie C, et al. Clinical and molecular consequences of microsatellite instability in human cancers. *Bull Cancer*. 2008;95:121–32.
46. Collura A, Lefevre JH, Svrcek M, Tougeron D, Zaanen A, Duval A. Microsatellite instability and cancer: from genomic instability to personalized medicine. *Med Sci (Paris)*. 2019;35:535–43.
47. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, Blosser RL, Fan H, Wang H, Luber BS, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov*. 2015;5:43–51.
48. Urbanski LM, Leclair N, Anczukow O. Alternative-splicing defects in cancer: splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA*. 2018;9:e1476.
49. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011;44:53–7.
50. Yip BH, Steeples V, Repapi E, Armstrong RN, Llorian M, Roy S, Shaw J, Dolatshad H, Taylor S, Verma A, et al. The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. *J Clin Invest*. 2017;127:2206–21.
51. Tsai YH, VanDussen KL, Sawey ET, Wade AW, Kasper C, Rakshit S, Bhatt RG, Stoeck A, Maillard I, Crawford HC, et al. ADAM10 regulates Notch function in intestinal stem cells of mice. *Gastroenterology*. 2014;147(822–834):e813.
52. Garcia Del Arco A, Edgar BA, Erhardt S. In vivo analysis of centromeric proteins reveals a stem cell-specific asymmetry and an essential role in differentiated, non-proliferating cells. *Cell Rep*. 2018;22:1982–93.
53. Lee LJ, Papadopoulos D, Jewer M, Del Rincon S, Topisirovic I, Lawrence MG, Postovit LM. Cancer plasticity: the role of mRNA translation. *Trends Cancer*. 2021;7:134–45.
54. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*. 2014;30:1015–6.
55. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
56. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31:213–9.

57. Corvelo A, Hallegger M, Smith CW, Eyras E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol*. 2010;6:e1001016.
58. Bauer K, Nelius N, Reuschenbach M, Koch M, Weitz J, Steinert G, Kopitz J, Beckhove P, Tariverdian M, von Knebel DM, Kloor M. T cell responses against microsatellite instability-induced frameshift peptides and influence of regulatory T cells in colorectal cancer. *Cancer Immunol Immunother*. 2013;62:27–37.
59. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U. HLA typing from RNA-Seq sequence reads. *Genome Med*. 2012;4:102.
60. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, Liu CJ, Zhao S, Feng YY, Graubert AP, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res*. 2020;8:409–20.
61. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res*. 2008;36:W509–512.
62. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020;48:W449–54.
63. Wang G, Wen WX, Mead AJ, Roy A, Psaila B, Thongjuea S. Processing single-cell RNA-seq datasets using SingCellaR. *STAR Protoc*. 2022;3:101266.
64. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19:15.
65. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2018;37(1):38–44.
66. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc*. 2020;15:1484–506.
67. Dai M, Pei X, Wang XJ. Accurate and fast cell marker gene identification with COSG. *Brief Bioinform*. 2022;23:bbab579.
68. Jonchère V, Montémont H, Letourneur Q, Collura A, Duval A. Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation. *EGA*. 2024. <https://ega-archive.org/studies/EGAS00001004863>.
69. Jonchère V, Montémont H, Letourneur Q, Collura A, Duval A. Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation. *GitHub*. 2024. https://github.com/CRSA-MSI/pipelines_and_scripts_Jonchre_Montemont_et_al_2024.
70. Jonchère V, Montémont H, Letourneur Q, Collura A, Duval A. Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation. 2024. *Zenodo*. <https://doi.org/10.5281/zenodo.12706170>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.