UNIVERSITÀ DI TORINO

Doctoral School in Life and Health Sciences

PhD Program in Complex System for Life Sciences XXXV Cycle

Genomic characterization, profiles of new antigens and biological impact of inactivation of Mismatch Repair genes in tumor cells

Author: Giuseppe Rospo

Tutors: Prof. Alberto Bardelli
Prof. Enzo Medico

Coordinator: Prof. Michele De Bortoli

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

October 2022

A Giovanni.

Per la tua generosità,
la tua curiosità,
ed il tuo rigore.

Table of Contents

Abstract 7

Introduction 9

- Microsatellite stable and unstable colorectal cancers 9
- Biological and mutational features of mismatch repair defective cancers 10
- Immunological characteristics of MMRd cancers 12
- Computational resources for neoantigen identification 16
- Non-canonical neoantigens 18
- Deciphering the non-canonical neoantigen landscape of MMRd cancers 20

Aim of the study 23

- Exploring the genomic and the biological characteristics of neoantigens in MMRd cancers 23

Materials and methods 25

- Cell line 25
- Gene editing 25
- Animal studies 25
- Hybridomas and antibodies 26
- Whole Exome Sequencing analysis 26
- Mutational signature analysis 27
- Immune-peptidomic workflow 27
- Whole Genome Sequencing analysis 29
- RNA Sequencing analysis 30
- Generation of specific peptide database for mass-spectrometry data search 31
- MHC-I associated peptide annotation 32

Results 33

- Mutational and neoantigen landscape of MMRp and MMRd CRC cells 33
- Mutational signature profiles of MMRp and MMRd CRC cell models 36
- Characterization of edited alterations in MMR-proficient and -deficient CRC cells 39
- Identification of MHC class I associated peptides 42
- Identification of edited peptides in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ tumor cells 45
- MAP classification in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ murine CRC cell line 47

Discussion 53

Conclusions and future perspectives 57

References 59

Abstract

Therapies based on immune checkpoint blockades (ICBs) are highly effective in patients affected by colorectal cancer (CRC) with mismatch repair deficiency (MMRd). These tumors carry a high number of mutations, which are assumed to be translated into a wide set of neoepitopes. A systematic classification of the neoantigen landscape in CRC carrying diverse MMR damages is lacking. Moreover, analyses based on mass-spectrometry peptidomics demonstrated the existence of MHC class I associated peptides (MAPs) originating from non-coding DNA regions. Based on these premises we investigated DNA genomic regions responsible for generating MMRd-induced peptides.

We exploited whole exome sequencing (WES) data of CT26 mouse model in which the MMR genes *Mlh1, Msh2, Msh6* and *Pms2* were genetically inactivated. A well-established computational pipeline was employed to characterize the mutational and the neoantigen landscape of those CRC cells. Further to this, CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ were inoculated in immunocompromised and - competent mice, and whole genome sequencing (WGS) and RNA sequencing (RNAseq) data were generated. First, peptide databases were built from transcriptomes of MMR proficient (MMRp) and MMRd cells. A database of peptides lost after injection in immunocompetent mice was generated from RNAseq data since those sequences were assumed to be edited by the immune system. Liquid chromatography-mass spectrometry (LC-MS) and matched transcriptome and whole genome databases were ultimately employed to identify the DNA regions from which the immune-edited MAPs originated.

The inactivation of *Mlh1, Msh2, Msh6* and *Pms2* in CT26 leads to the acquisition of several mutations during time as compared to the MMRp CT26. In addition, WGS analyses revealed an unbalanced distribution of immune edited alterations across the genome of *Mlh1*$^{-/-}$ cells grown in immunocompetent mice. The integrated computational and LC-MS analyses also revealed that immune edited MAPs originated mainly from atypical translational events in both MMRp and MMRd models. Moreover, CT26 *Mlh1*$^{-/-}$ showed a strikingly different repertoire of mutant MAPs targeted by the immune system, mainly derived from untranslated regions (UTRs) and out- of-frame translation of coding regions.

Our results suggest that MMRd tumors generate a significantly higher number of neoantigens, compared to MMRp CRC, mainly classified as non-canonical mutated peptides that bind the MHC class I. These results reveal the importance of evaluating the diversity of neoepitope repertoire in MMRd tumors to identify novel neoantigens as therapeutic targets and further understand the reason why these tumors are highly responsive to ICB treatments.

Introduction

Microsatellite stable and unstable colorectal cancers

Latest global cancer statistics revealed that colorectal cancer (CRC) is currently the second leading cause of cancer death and the third most commonly diagnosed cancer worldwide (1). Sixty-five percent of CRC patients are estimated to survive 5 years after being diagnosed with cancer. The same expectation is reduced to 15% when considering the metastatic stage (2). A molecular profiling could facilitate the examination of therapeutic options that may be available for treating metastatic CRC (mCRC) patients based on tumor subtypes (3). The vast majority of CRCs are classified as MMR proficient (MMRp) and the length of microsatellites is stable (MSS) over time. A noteworthy proportion of colorectal tumors is classified as mismatch repair deficient (MMRd) owing to methylation or mutations in *MLH1, MSH2, MSH6* and *PMS2* genes. They often display a shifting length of microsatellites and are classified as microsatellite unstable (MSI) tumors. MMRd tumors account for 15% of all CRCs and 5% of mCRCs (4). In patients with MSI or MMRd tumors, treatment based on immune checkpoint blockade (ICB) extends survival significantly more than conventional therapeutic options. As a matter of fact, in 2020 FDA approved the immune checkpoint inhibitor pembrolizumab for first-line therapy of MSI/MMRd mCRC patients (3).

In the MSS subtype, tumor progression is driven by the so-called "chromosomal instability" characterized by acquisition or loss of an entire or part of chromosome(s) that occurs in association with genomic alterations in proto-oncogenes or tumor suppressor genes. On the contrary, in MSI tumors the chromosomal integrity is not affected by the genomic instability. The progression of disease is instead promoted by the accumulation of insertions or deletions in short nucleotides repetitive regions (microsatellites) owing to non-functional MMR (5). MSI tumors exhibit peculiar genetic and clinical-pathological features since they are more frequently located in the right side of the colon, and they show mucinous features and poor histological differentiation. Moreover, they are characterized by a great amount of tumor mutations – not only limited to single mismatches but rather also short insertions and deletions - and high number of tumor-infiltrating lymphocytes, mainly represented by CD8$^+$ T-cells (6-8). About one third of MSI tumors is represented by a hereditary mutational mechanism that affects one of the products of the MMR system, that is the nonpolyposis colorectal cancer (HNPCC) syndrome, best known as Lynch syndrome (5). Sporadic forms of MSI CRC are the majority of MSI tumors caused by epigenetic hypermethylation and the consequent inactivation of *MLH1* gene. At the molecular level, MSI tumors often carry BRAF mutations, while KRAS alterations are less frequent (5). Furthermore, the incidence of APC and p53 mutational variations is higher in MSS than MSI tumors (9) (Figure 1). In addition, MSI mCRCs are often resistant to common cytotoxic agents (10). In addition to this - while MSS mCRCs exhibit frequent primary resistance to ICBs - MSI mCRCs are greatly sensitive to immune checkpoint inhibitors (3, 11-13).

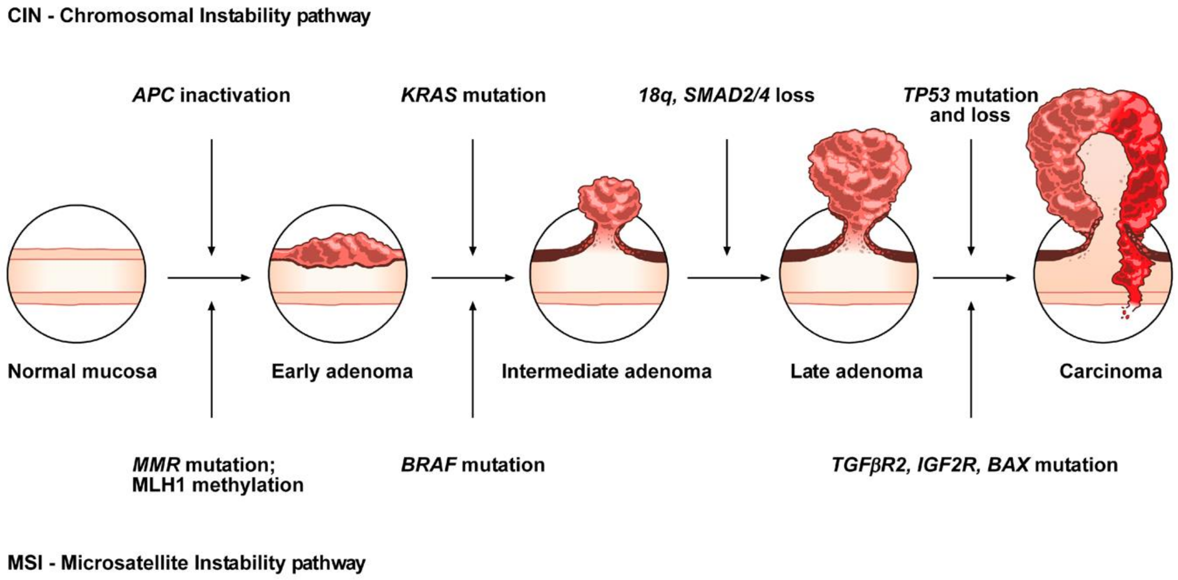

Figure 1 CRC adenoma-carcinoma sequence as a multistep mutational pathway (5).

Biological and mutational features of mismatch repair defective cancers

The MMR machinery is composed of several multi protein complexes which are able to detect and correct erroneous substitutions, such as single nucleotide variants (SNVs), insertions and deletions (indels) following DNA replication (14). The four components of the MMR system are MuL homolog 1 (MLH1), PMS1 homolog 2 (PMS2), MutS homolog 2 (MSH2) and MutS homolog 6 (MSH6). To guarantee the efficacy of the entire system these four molecules act as heterodimers (Figure 2): MSH2 and MSH6 compose the MutSa complex; MLH1 and PMS2 form the MutLa complex. Both complexes are capable of recognizing base-base mismatches and small indels. In addition to these, the MutSbeta heterodimer - composed by MSH2 and MutS homolog 3 (MSH3) - detects and corrects large indels (4).

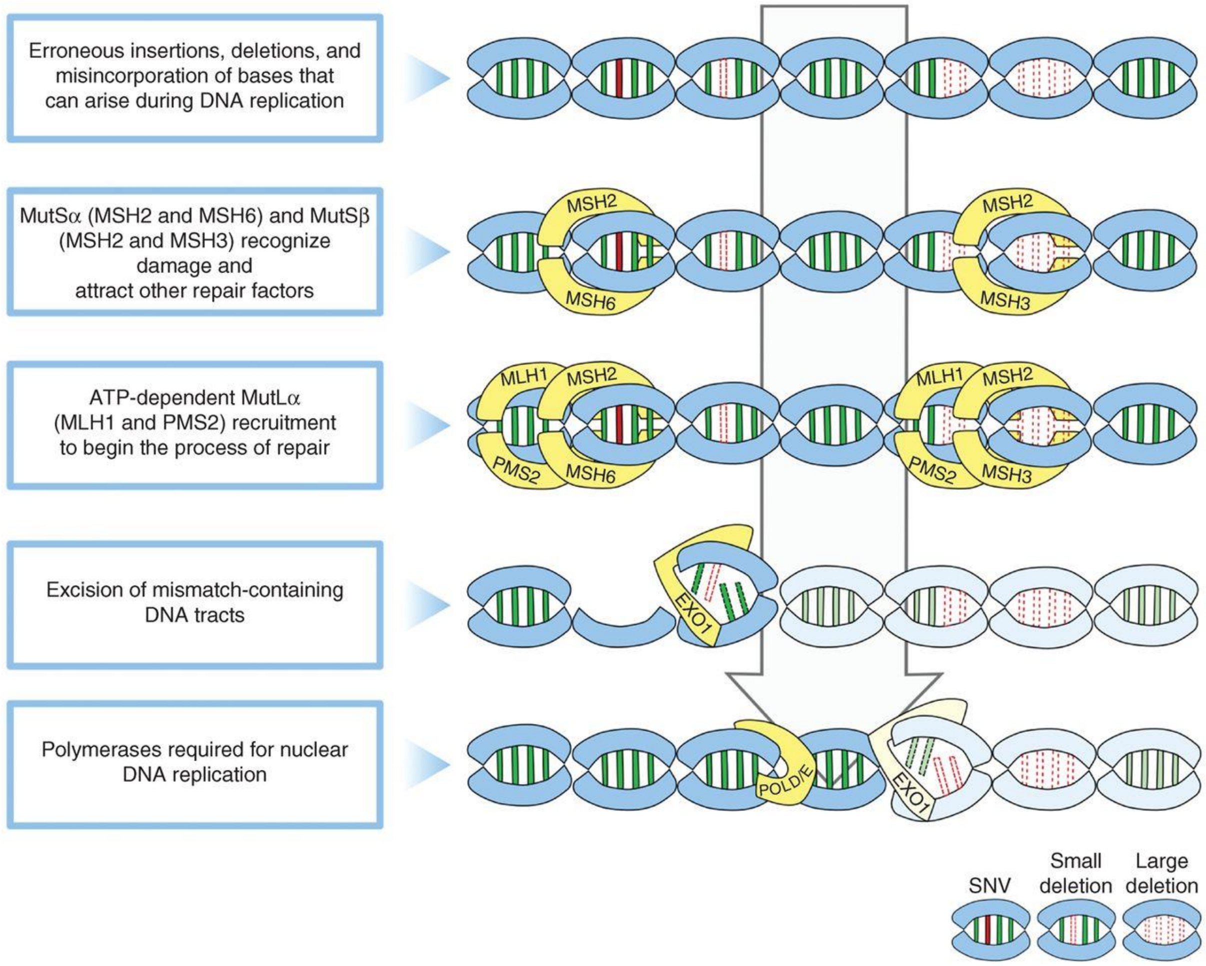
Erroneous insertions, deletions, and misincorporation of bases that can arise during DNA replication
MutSα (MSH2 and MSH6) and MutSβ (MSH2 and MSH3) recognize damage and attract other repair factors
ATP-dependent MutLα (MLH1 and PMS2) recruitment to begin the process of repair
Excision of mismatch-containing DNA tracts
Polymerases required for nuclear DNA replication
MSH2
MSH6
MSH2
MSH3
MLH1
MSH2
PMS2
MSH6
MLH1
MSH2
PMS2
MSH3
EXO1
POLD/E
EXO1
SNV
Small deletion
Large deletion

Figure 2 Molecular products of mismatch repair machinery (4).

Defects affecting one or more MMR products lead to DNA repair loss of function and contribute to carcinogenesis and microsatellite instability (5). Both genetic and epigenetic events are involved in the onset of MMRd status and the emergence of MSI. The inheritance of mono-allelic alteration in one MMR gene contribute to cancer disorders such as Lynch syndrome. Germline biallelic inactivation of MMR genes causes the so-called constitutional MMR deficiency (cMMRd), a rare disease that is associated with early CRC onset and pediatric cancers (4). However, only 3% of all CRCs - carrying microsatellite instability - emerge in the context of HNPCC. The majority of MSI CRCs develop due to somatic mutations in MMR genes or epigenetic downregulation of MLH1 expression (15).

The inefficient DNA repair system leads MMRd tumors to accumulate a 10-fold increase of unsolved alterations across the genome compared to MMRp tumors (6, 16, 17). Indeed, defects of MMR machinery leads to the emergence of genetic variations such as SNVs, which affect an individual amino acid, and small indels, which can lead to the generation of frameshift variants (new amino acid frame). SNVs and indels alterations are usually identified combining Next Generation Sequencing (NGS) sophisticated bioinformatic tools (4, 6).

Notably, recent studies demonstrated that MMRd induced mutations follow specific patterns of DNA alterations (18). Mutational patterns driven by deficiency in MMR pathway can be identified looking at mutational signatures: they are able to unveil the different etiologies of mutational processes, including DNA damage, repair and/or replication mechanisms. This procedure classifies each genomic alteration by looking at the base immediately 5′ before the somatic mutation and the base immediately 3′ after the somatic mutation. Thus, from each of the six types of somatic variants (C>A, C>G, C>T, T>A, T>C and T>G) result 16 different substitutions that generate a total of 96 unique mutation types (19). MMRd tumors show specific mutational patterns composed by the enrichment of C>T and T>C mutation types. They are also characterized by other peculiar patterns: double base substitution and small insertion deletion (18). Moreover, the combination of distinct mutational processes may generate different mutational signatures, such as MMRd and Polymerase Epsilon (POLE)/Polymerase Delta 1 (POLD1) mutant tumors (20, 21).

Immunological characteristics of MMRd cancers

Mutations induced by MMR damages directly affect the mutational landscape of genomic DNA, in terms of both quantity and quality values. The type of DNA repair defects occurring in CRCs can affect the type of mutations. As a matter of fact, MSI tumors display a higher number of frameshift indels than *POLE* mutant lesions that, on the contrary, show many SNVs (6). Those alterations, if transcribed and translated, can be presented as peptides by the major histocompatibility complex (MHC) class I and II to the and trigger adaptive immunity (8, 22). At protein level, non-synonymous SNVs drive the emergence of new epitopes that differ from the *wild type* sequence only for one amino acid. On the contrary, frameshift indels lead to the generation of new peptides that vary greatly from their *wild type* counterpart. In addition to this, indel mutations that cause a frameshift produce new open reading frames that potentially give rise to a large amount of neoantigenic peptides. Then, indel variants can both increase the number of mutant peptides and reduce susceptibility to self-tolerance mechanisms (23).

The hypermutation status of MSI tumors is associated with increased responsiveness to immune-based therapies, such as ICBs (11, 12, 24). Indeed, MSI tumors show a high tumor mutational burden (TMB), which varies significantly across cancer types (25) but also inside the same histology (6). Furthermore, the genomic landscape driven by MMRd uniquely contributes to the quality of the neoantigen profiles since the generation of frameshift indels augments the number of putative neoepitopes per each event and, consequently, the generation of immunogenic neoantigens. In addition to this, also the high number of non-synonymous SNVs positively contribute to neoantigen load in CRCs (26). Therefore, tumors with a high number of neoantigens show increased response to ICBs. On the contrary, tumors with fewer mutations - and neoantigens - are more likely to be unresponsive to immunotherapy (27).

The association between such peculiar biological and clinical features of MMRd tumors lies in the immunological properties that these tumors unveil. Indeed, high levels of neoantigens were positively associated with overall lymphocytic infiltration, tumor-infiltrating lymphocytes (TILs), memory T cells, and colorectal cancer–specific survival (26). Moreover, the presence of higher TILs in CRCs has long been recognized as evidence of microsatellite instability (28, 29) (Figure 3). Also, Galon et colleagues proved that the immune repertoire - the type, density, and localization of infiltrating lymphocytes in the center and the invasive margins - of colorectal cancers is an independent prognostic factor and positive predictors of better survival (30). During the last decade the same group defined and improved the concept of "immunoscore": a classification criterion that is based on the localization and the amount of CD3$^+$ and CD8$^+$ T cell subpopulations in the tumor microenvironment (31). Moreover, ESMO guidelines have included the immunoscore for the staging of CRC since 2020 (32). Of relevance, the immunoscore value is able to estimate the risk of recurrence in CRCs regardless of MMR status (33).

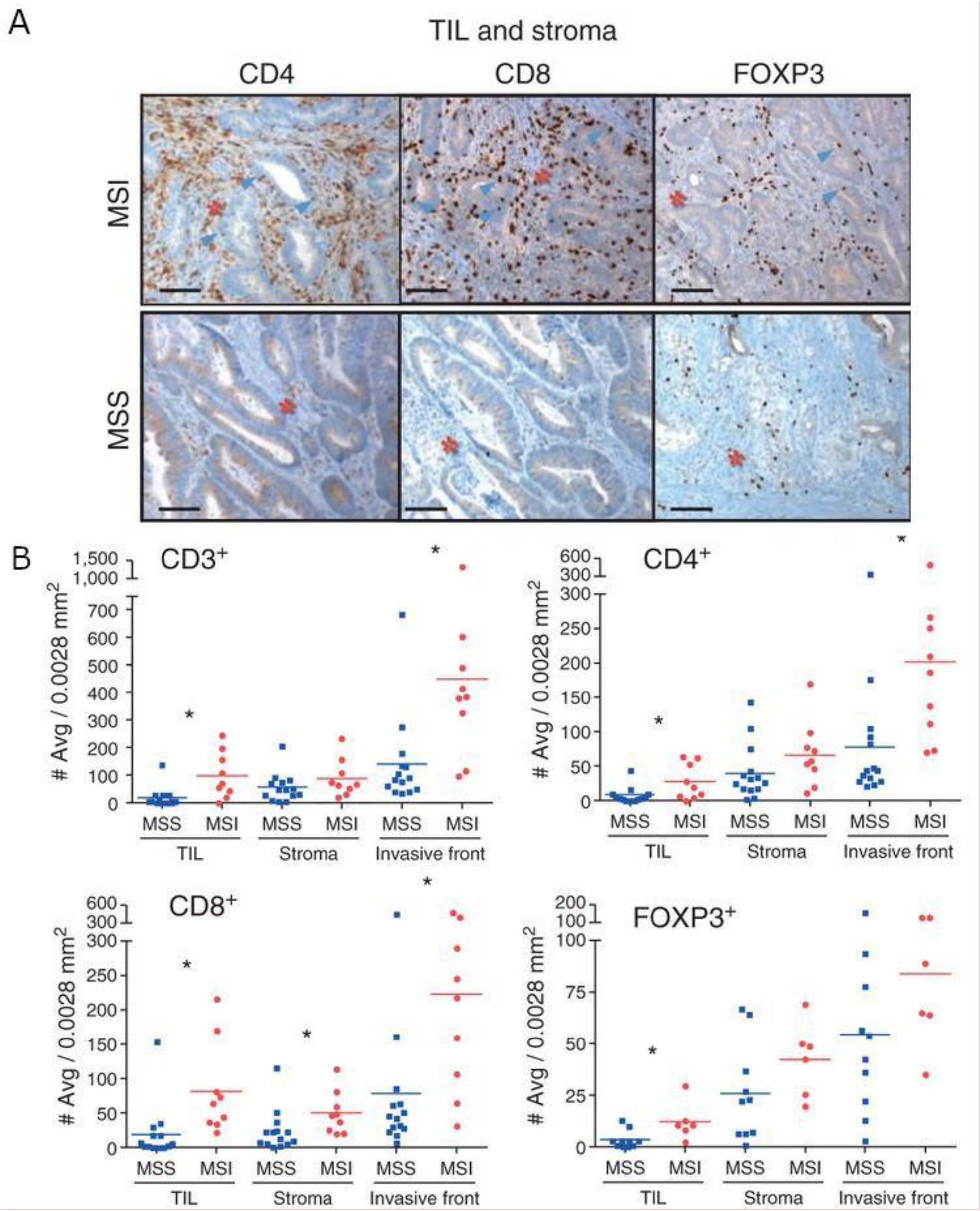

Figure 3 Spatial distribution of tumor infiltrating lymphocyte in MSS and MSI CRC **A)** Immunohistochemical staining of CD4$^+$, CD8$^+$, and FOXP3$^+$ cell infiltration (red stars and blue arrows indicate the tumor stroma and tumor epithelium-infiltrating immune cells, respectively). **B)** Cell density quantification (adapted from (29)).

Tumor microenvironment of MSI tumors and the response to ICBs are highly conditioned by the TMB and then the neoantigen landscape of these tumor types. Germano and colleagues demonstrated that CRC mouse models acquire a wide and dynamic mutational spectrum after being inactivated by the *Mlh1* gene (8). They analyzed MMRd and MMRp tumors inoculated in immunocompetent and -compromised mice and noted a CD8$^+$ T-cell directed response in *Mlh1*$^{-/-}$ tumors.

Moreover, MMRd tumors triggered increased levels of T-cell rearrangements (TCR) in the blood of mice as compared to MMRp cancer cells (8). In the same line of evidence, several studies highlighted the correlations between genomics, immune cells infiltration and better response to ICB treatments (26, 34, 35).

These findings lead MSI tumors to be the proper target for a successful treatment with ICBs (36). The capacity of ICBs to induce adaptive immune responses is fascinating. The interplay between the immune system and cancer cells is indeed negatively affected through the upregulation of specific immune receptors present on tumor surfaces which magnify the immune-suppressive environment caused by cancer cells (37). ICBs are monoclonal antibodies that target specific immune checkpoint products and reactivate T-cells antitumor activity. Following these considerations, immunotherapy treatments result highly effective in mCRC MSI patients who failed previous lines of treatments. Patients showed an outstanding 40% of objective response rate (ORR) with a 90% disease control rate (DCR) - as compared to 0% ORR and 11% DCR in patients carrying MSS tumors (11). Moreover, ICBs demonstrated to be very effective also in MMRd non-colorectal tumors (24, 38-40) (Figure 4).

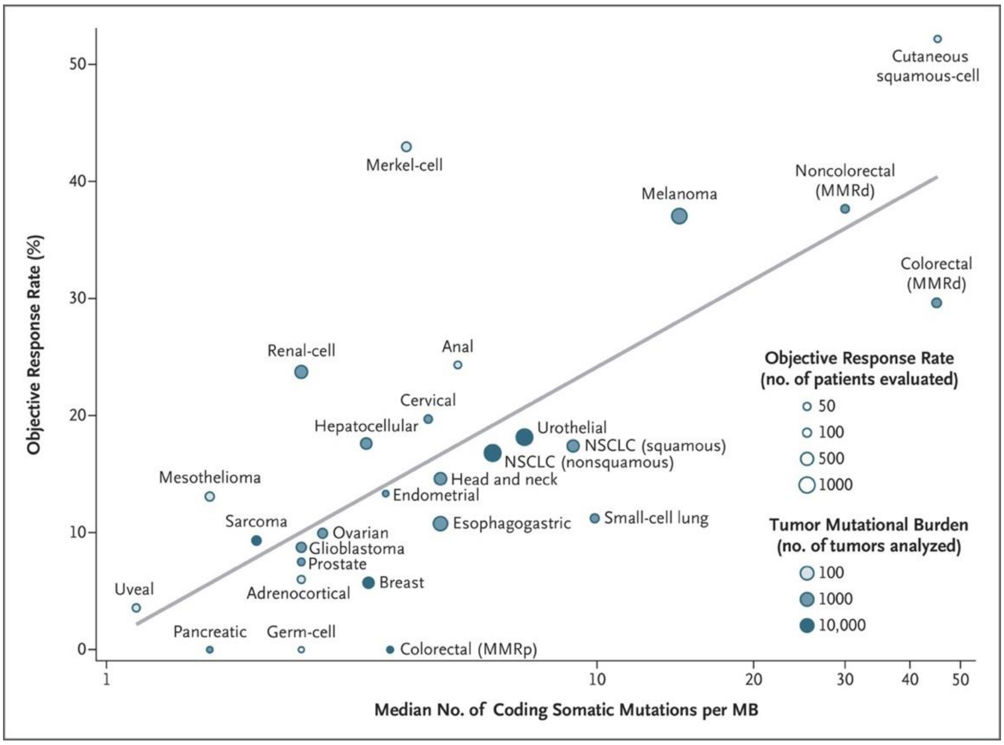

Figure 4 Correlation between tumor mutational burden and objective response rate upon anti–PD-1 or anti–PD-L1 treatment in several tumor types (38).

Following such remarkable clinical results, in 2017 the FDA approved pembrolizumab for the treatment on any MSI solid tumors. Further to this, very recently, Cercek and colleagues reported a clinical complete response in all 12 patients with rectal cancers treated with dostarlimab (PD-1 inhibitor). Remarkably, none of the patients had received chemoradiotherapy or undergone surgery, and they have undergone at least 6 months of follow-up (41).

Computational resources for neoantigen identification

A comprehensive neoantigen characterization can be performed by using NGS data. Indeed, several advanced bioinformatic pipelines are available to identify immune activating neoantigens starting from genomic data (42-45). Vast majority of bioinformatic tools present as core function the prediction of the affinity binding between a peptide sequence and the MHC class I or II (46, 47). Further to this, recent approaches integrate multiple features to refine the quality of results such as gene expression data, variant allele frequency and analysis of clonality. Neoantigen prediction software can be classified in three main categories: a) ready-to-use tools - which appear like a black boxes ready to use but hardly or not at all customizable; b) recommended pipelines - which usually are built as multi-step processes with default parameters and tools but allowing the possibility to perform some changes; c) algorithms and tools that specifically perform HLA binding prediction - they are not always ready to use but moderately adaptable to custom pipelines. The very first step for the prediction of tumor neoantigens is performed by the identification of DNA alterations (46, 47). Commonly the mutational characterization is performed by exploiting WES data from paired normal and tumor samples. The analysis includes single base mismatches, aberrations derived from insertions and deletions, splice variants, gene fusions and other genomic alterations that could generate non-self-peptides (47). As reported before, MMRd tumors are highly enriched of SNVs and frameshift indels, therefore these alterations are deeply characterized when MMRd cancer data undergo bioinformatic analysis (4). Next, mutated sequences are filtered according to transcript expression - if RNA sequencing (RNAseq) data are available - and then mutated sequences are translated and properly processed before being analyzed for MHC affinity binding prediction (4) (Figure 5).

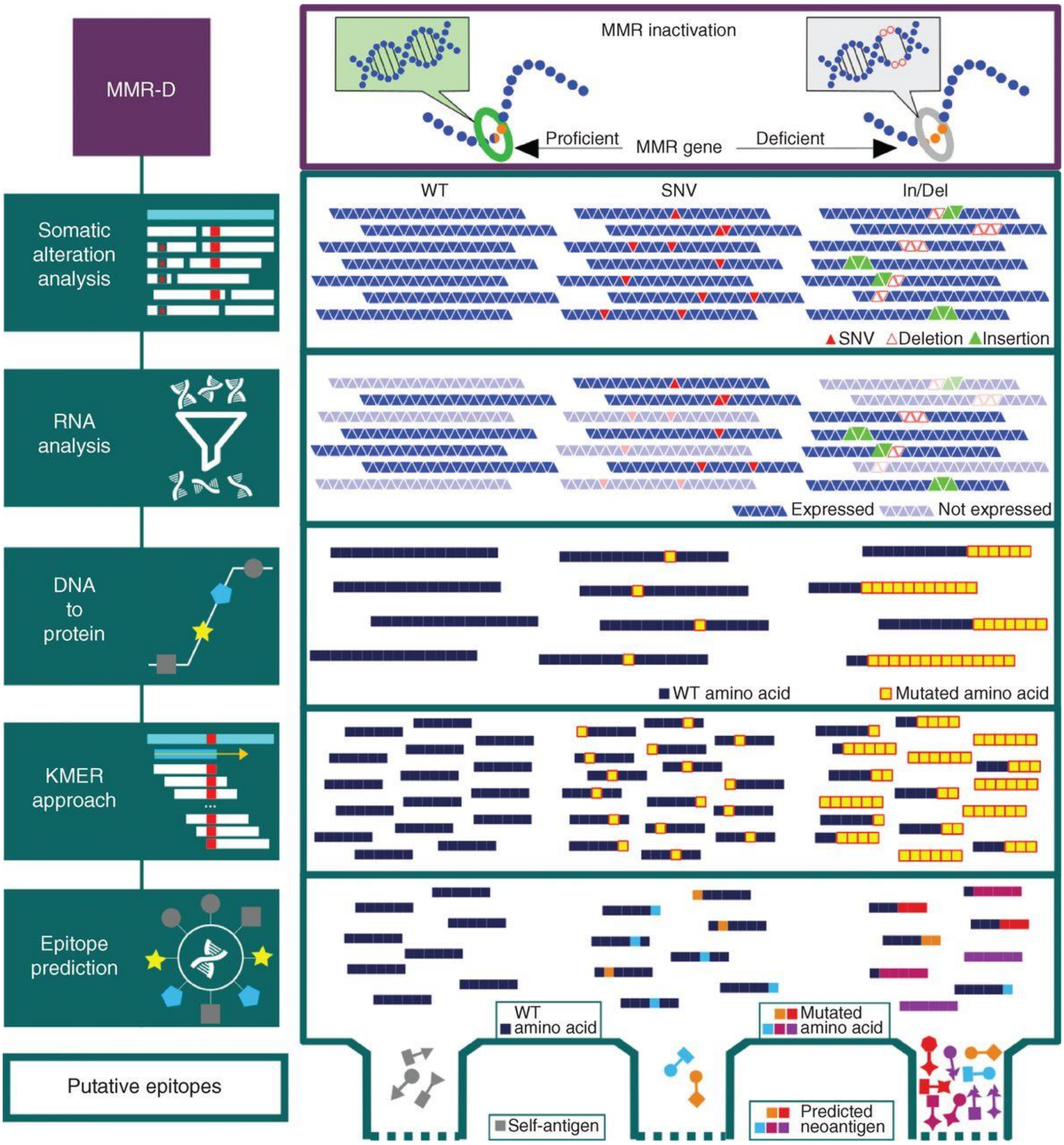

Figure 5 *In silico* neoantigen prediction pipeline (4).

Further to this - and if not known a priori - an accurate identification of the HLA allele is necessary to properly predict the affinity binding between MHC and neopeptides. The standard HLA typing procedure is performed using serology- or PCR-based methods (48). However, both serological and molecular genotyping, which are time-consuming, laborious, and expensive, do not meet the increasing requests from clinicians and researchers. With the advent of NGS, several computational methods are now available that allow HLA genotyping by inferring WGS, WES or RNAseq data

(48). Moreover, it is worth mentioning that even though multiple studies highlighted the role of MHC class II restricted neoantigens for therapeutic purposes, most of the pipelines focus on MHC-I neoepitopes due the higher accuracy of available predictors and the greater quantity of MHC-I ligand entries in database such The Immune Epitope Database (IEDB) (49).

Despite a huge amount of information can be retrieved by using these advanced computational approaches - that is identifying tumor mutations, computing HLA genotyping and predicting HLA-peptide binding affinity in a high throughput fashion - they still lack accurate sensitivity. For this reason, validation of cancer neoantigens is needed in clinical - but also in research – practice (50). First, predicted neoantigens should be validated to bind the MHC. Indeed, only a small fraction of mutant putative peptides will be processed and presented on the cell surface by the MHC. Next, the immunogenic properties of those peptides should be tested, since a reduced number of MHC associated neoantigens will be recognized by a T-cell receptor (TCR)-bearing T-cells. Previous analyses demonstrated that only 1% of predicted neoantigens bind MHC while half of these are recognized by T-cells and only one third are regularly processed allowing target cell killing (50).

In conclusion, both practices - high-throughput computing of neoantigen prediction and validation assay - present benefits and limitations. Overall, the computational approaches previously described are pivotal to predict immunogenic neoantigens; however new tools to obtain more sensitive prediction and more rapid immunogenicity validations are needed.

Non-canonical neoantigens

Computational analyses of cancer immunogenicity are currently based on exome data or on a limited number of target genes (custom panels) (51-53). Indeed, neoantigen prediction analysis is based on the identification of somatic non-synonymous mutations in canonical annotated protein regions by WES and the prediction of the binding affinity between MHC and mutant peptides. Although the contribution of neoantigens in deciphering the immunogenic features of these tumors has been well described (4, 26, 36), the extent to which the non-coding portions of the genome affects the immunogenicity of MMRd tumors is largely unknown. Indeed, several lines of evidence suggest that a variety of non- coding regions can contribute to the repertoire of tumor antigens (54-57). They include novel or unannotated open reading frames (58), retained introns (59), long noncoding RNAs (60), untranslated regions (UTRs) (61, 62), junctions and intergenic regions (55, 63) (Figure 6).

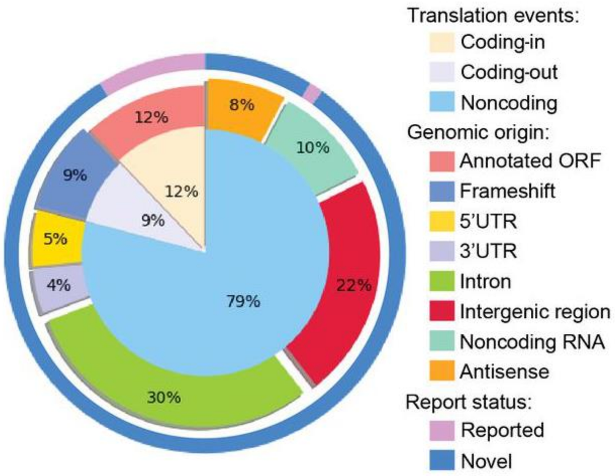

Figure 6 Source of unmutated tumor specific antigens across the entire genome (57).

As matter of fact, Laumont et colleagues first reported that a relevant subset of MHC class I associated peptides (MAPs) derived from non-canonical reading frame (63); then with their proteogenomic approach reported that noncoding regions are the main source of targetable tumor-specific antigens (TSA): they identified 40 TSAs, about 90% of which were derived from allegedly noncoding regions (55). Interestingly, MAPs originating from non-coding portions of the genome were shown to be potential immunogenic targets of T-lymphocytes (55, 64). MAPs can also derive from a variety of genetic and epigenetic changes leading to the transcription and translation of genomic sequences normally not expressed in cells or from non-canonical open reading frames that emerge in tumor cells (65, 66). Moreover, a recent study demonstrated that non-canonical open reading frames (ORF) encode functional proteins essential for cancer cell survival. The authors reported that 50 out of about 500 candidates from noncanonical ORF datasets induced viability defects when knocked out in human cancer cell lines (67).

Given the potential relevance of non-coding MAPs and mutated MAPs (mMAPs) in the immunogenic properties of MMRd tumors, we thought that further investigation at immunopeptidome level was necessary, to systematically analyzed how the immune system could perturb the canonical and non-canonical antigen repertoire of MMRd and MMRp tumors.

Deciphering the non-canonical neoantigen landscape of MMRd cancers

The main routine practices currently used to detect neoantigens relies on identification of tumor somatic mutations by using exome sequencing data, followed by an in-silico prediction analysis of candidate mutated peptides. The accuracy of peptide-MHC binding prediction algorithms - which are often poorly accurate - highly affect the success of this method. Moreover, given the potential immunogenic features of neoantigens generated from non-coding regions, these workflows present a major limitation due to the impaired ability to analyze other genomic regions than the canonical coding portions. Recent advances in mass spectrometry allow the analysis of the HLA peptidome - which consists of the set of HLA class I bound peptides expressed by a specific cell - in great resolution (68, 69). This strategy combines NGS data and mass spectrometry analysis and allows to define the immunological signature that can lead to the identification by the immune system cells. Briefly, in parallel to sequencing data analyses, an immunoaffinity purification of the HLA molecules from the same cells is performed and then a tandem mass spectrometry (MS) analysis of HLA peptides is generated. The MS spectra are analyzed by specific tools, such as MaxQuant (70), and matched against a specific peptide dataset, which should include the mutant variants inferred by previous sequencing data analysis (69) (Figure 7).

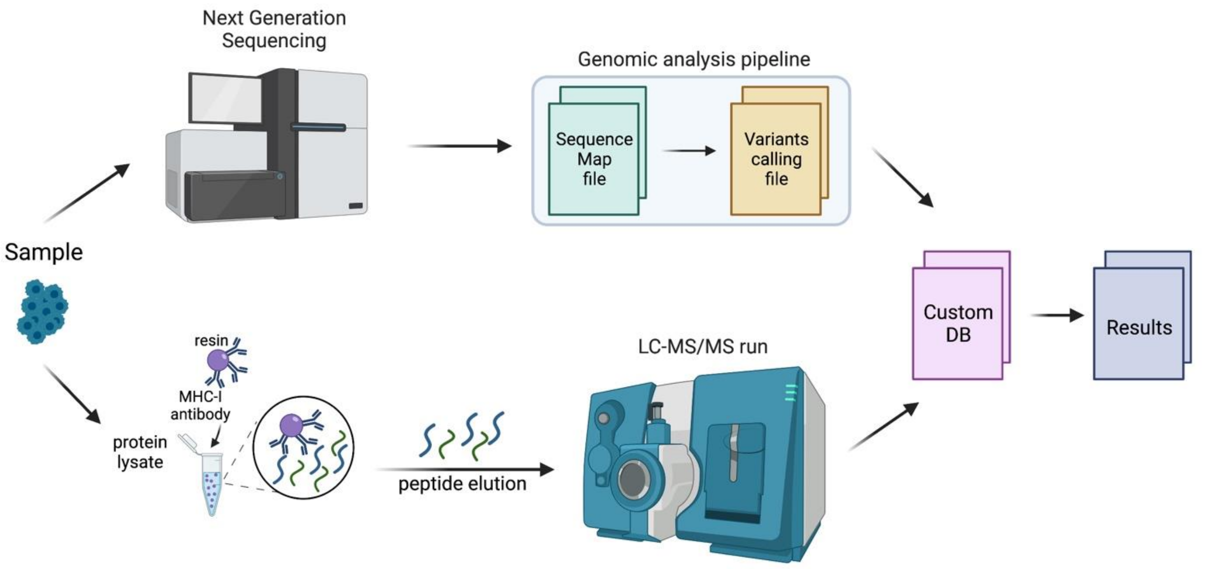

Figure 7 Example of immune-peptidomic pipeline. Tumor sample undergoes NGS and mass-spectrometry analysis. The final results are generated by matching NGS-derived database and spectra data.

In 2016, Kalaora and colleagues applied the immune-peptidomic pipeline to a melanoma patient, and they identified two mutant peptides derived from WES analysis and validated one of them for reactivity with autologous bulk tumor infiltrating lymphocytes (69). To enlarge the possibility of detecting unconventional neoantigens generated from all genomic regions other studies performed MS-based peptide sequencing by matching customized databases which included all-frames translation of genomic or transcriptome sequences (55, 63). In this way, immunopeptidome studies can detect peptides coded by all reading frames from every genomic region. Laumont and colleagues generated an all-six frame translation peptide database by using the transcriptome of human B-lymphoblastoid cell lines. This database was used to identify MAPs by matching the high-throughput MS sequencing data. Integrating transcriptome and proteomic data they classified as cryptic about 10% of MAPs (63). In a subsequent study the same group revealed that most of the tumor specific antigens generate from non-coding regions (55). Although this strategy allows evaluating the peptides bound to the MHC-I, it is cell dependent and therefore it is still difficult to apply in a high-throughput fashion.

Recent studies unveiled novel approaches to investigate the genomic regions that are transcribed and translated (71, 72). Ingolia and colleagues designed ribosome profiling, a method to systematically characterize translated sequences. This strategy relies on deep sequencing of ribosome protected messenger RNA (mRNA) fragments and permits to unveil the translation activity in a cell with single-nucleotide resolution (71, 72). Indeed, evaluating the density of sequences protected by translating ribosomes - which are called ribosome footprints (RF) - gives a measure of the protein synthesis rate. Moreover, the identification of genomic coding regions is simplified by the start and stop codons presence in RFs, as compared to standard RNAseq which only provides an estimation of the transcript borders. Several studies that analyzed ribosome-profiling data highlighted that some predicted noncoding regions of the transcriptome were actively translated (62, 72, 73). As example, Chen and colleagues recently demonstrated that 240 non-canonical peptides derived from upstream open reading frames located in the 5’UTR and long non-coding RNAs of extragenic DNA were presented by the HLA of human tumor cell lines (62). Despite ribosome profiling being a further step to the exceptional advancement of NGS methods, the ability to calculate and predict with high sensitivity the binding affinity between HLA and peptides in a high throughput fashion is still lacking.

Aim of the study

Exploring the genomic and the biological characteristics of neoantigens in MMRd cancers

The use of immunotherapy based on checkpoint blockade produced outstanding results in MMRd colorectal tumors. Indeed, defects in the MMR machinery result in the accumulation of genomic alterations which are predicted to translate into a wide repertoire of neoepitopes. Interestingly, analysis based on mass-spectrometry peptidomics has demonstrated the existence of MAPs originating from all genomic regions.

Based on these premises we aimed to investigate the role of DNA repair genes, *Mlh1, Msh2, Msh6* and *Pms2*, to understand how damages in the MMR pathway could perturb the mutational landscape of CRC and how this could result beneficial in clinical practice. Moreover, given the potential relevance of non-coding MAPs and mMAPs in the immunogenic properties of MMRd tumors our aim is to analyze how the immune system could perturb the canonical and non-canonical antigen repertoire of MMRd and MMRp tumors.

Briefly, the goal will be to characterize the specific DNA variants triggered by inactivation of the above-mentioned genes as function of putative neoantigens in terms of quality and quantity. In addition, we will characterize how the neoantigens generated in MMRp and MMRd CRC model are edited by an immunocompetent host. Considering the challenges in functionally characterizing these aspects in human models, we exploited a well characterized isogenic murine CRC model in which we previously perturbed MMR proficiency through gene knock-out with the CRISPR/Cas9 technology.

Since 99% of cancer mutations are in non-coding regions, we postulate that non-coding DNA could be a source of novel MAPs and contribute to the high immunogenic impact of MSI tumors when treated with ICB.

Materials and methods

Cell line

CT26 is a chemically induced colon carcinoma derived from BALB/c mice; CT26 cells were cultured in RPMI 1640, 10% FBS, 1% glutamine, 1% penicillin and streptomycin (Sigma Aldrich). Cells were regularly checked for mycoplasma contamination and before performing the genome editing experiments, they were injected into matched syngeneic mice to ensure cell tumorigenicity. After tumor formation, we established again *in vitro* cell cultures. All cells underwent WGS.

Gene editing

To knockout the *Mlh1, Msh2, Msh6* and *Pms2* genes, we used the genome editing one vector system (lentiCRISPR-v2) (Addgene #52961) as previously reported (8). Briefly, sgRNAs were designed using the CRISPR tool (http://crispr.mit.edu) to minimize potential off-target effects. For transient expression of CRISPR-Cas9 system, we transfected cells with lentiCRISPR-v2 vector plasmid (same guides as previously described) (8). Transfection was carried out using Lipofectamine 3000 (Life technologies) and Opti-MEM (Invitrogen), according to the manufacturer's instructions. After 48 hours CT26 cells were incubated with puromycin (Sigma Aldrich) for 2 days and subsequently single cell dilution was performed in 96-well plates. The absence of MLH1 and CAS9 was confirmed by western blot (8).

Animal studies

All animal procedures were approved by the Ethical Commission of the FIRC Institute of Molecular Oncology (IFOM) and by the Italian Ministry of Health, they were performed in accordance with institutional guidelines and international law and policies. Four-six weeks old female NOD-SCID and BALB/c mice were purchased from Charles River and were maintained in pathogen-free conditions in individually ventilated cages. CT26 cells were resuspended in PBS and injected (500000 cells per mouse) subcutaneously. When tumors reached 1200 mm^3 of volume they were explanted for subsequent analyses.

Hybridomas and antibodies

HB-79 (producing anti H-2Kd/H-2Dd mouse IgG2a) and HB-27 (producing anti H-2Ld mouse IgG2a) hybridoma cell lines were purchased from ATCC and grown in Iscove medium (Sigma) supplemented with 10% FBS. Hybridoma were then adapted to protein-free PFHM medium (Thermo) for expansion and conditioning. Once cells were dead, the medium containing immunoglobulins was centrifuged and filtered to be run on a MabSelect Sure (ProteinA) column (Cytiva) mounted on Akta Pure (Cytiva). IgGs were then eluted at acid pH and dialysed against physiologic storage buffer.

Whole Exome Sequencing analysis

Genomic DNA of cell lines was extracted using the ReliaPrep gDNA Tissue Miniprep System (Promega). The library preparation, exome capture and sequencing were performed by Integragen SA (Evry, France) on Illumina NovaSeq as paired-end 100 bp reads. Raw data provided by IntegraGen were analyzed at our institution using the bioinformatics pipeline previously published (74). Briefly, fastq files were aligned to the mouse reference mm10 using BWA-mem algorithm (75), and then polymerase chain reaction (PCR) duplicates were marked using MarkDuplicates in the Picard tools suite (76). On the resulting aligned files, we observed a median depth of 90x with 99% of the targeted region covered by at least one read. We noted that different sequencing led to high depth discrepancy among samples. For this reason, we applied a sampling approach to normalize raw data and reduce the depth discrepancy among the sequenced samples. Briefly, all the fastq files were downsampled by selecting the same number of starting reads (according to the samples that showed the smallest number of sequences). Firstly, the read names were extracted from the fastq files and randomly sorted. Then, 35 million reads were selected from each file and used as input with matched aligned files of Picard tools *FilterSamReads* (76) to generate downsampled cram files. This strategy was performed three times per sample. Bioinformatic modules previously developed (74) by our laboratory were used to identify count mutant alleles and consequently call SNVs and indels. Murine germline alterations were subtracted by using normal DNA of BALB/c mice previously sequenced at our institution. For calling mutations, we considered only positions present with a minimum depth of 5x and supported by at least 1% allelic frequency.

Tumor mutational burden was calculated as the number of variants per megabase considering those derived from coding regions. The prediction of neoantigens was performed using a bioinformatic pipeline we previously published (4, 6, 8, 77). In brief, NetMHC 4.0 software (44) was employed to analyze mutated peptides derived from SNV calls that were properly located in kmer composed by 8-11 amino acids. For frameshift indels, we applied the same approach considering every possible peptide generated by the new frame. Finally, haplotypes for murine samples were set to H2-Kd and H2-Dd (BALB/c background) and only peptides with predicted strong binding affinity (Rank < 0.5) were considered for further analysis.

Mutational signature analysis

Variant calling files previously generated were used to calculate the mutational signature profiles. The pyrimidine base of the Watson–Crick base pair were used to calculate the six substitution subtypes C>A, C>G, C>T, T>A, T>C, and T>G. To generate the 96 possible mutation types, information about the nucleotides immediately 5' and 3' to the mutation were retrieved from the reference genome and incorporated. Then, the 96 mutated trinucleotides were normalized according to the actual trinucleotide frequencies previously calculated on the mouse exome version mm10. The signature extraction and the analysis of mutations associated with post-replicative MMR deficiency was computed using *signal* (78). In brief, a tab-delimited in which each row corresponds to a single mutation in a particular sample was generated. Specifically, the information about sample name, chromosome, position, original base, and mutated base were selected. Finally, the variant file was used as input for the analysis performed by the web application (78).

Immune-peptidomic workflow

Six CT26 *Mlh1*$^{+/+}$ and six CT26 *Mlh1*$^{-/-}$ tumor masses were explanted from NOD-SCID mice and manually smashed with disposable micro tissue homogenizers in lysis buffer solution (0.25% sodium deoxycholate, 0.2 mM iodoacetamide, 1mM EDTA, 1:200 protease inhibitors cocktail, 1mM PMSF, 1% octyl-b-D glucopyranoside in PBS). Proteins were extracted for 1 hour at 4°C in continuous mixing, then samples were centrifuged at 30000 rpm for 1 hour at 4°C. Protein extracts in the supernatants were pre-cleared with 1 mL of protein A resin (GenScript) for 1 hour at 4°C in agitation, then dosed by BCA assay. Around 20 mg were used for reaction, and each experiment was performed 3 times.

Protein A resin was washed 3 times with PBS, then resuspended in PBS-Tween 0.01% and added with 5 mg of anti H-2Kd/H-2Dd or anti H-2Ld antibodies. Control samples without antibody were included. Resin and antibody were left with continuous mixing at 4° C overnight, then the unbound antibody was discarded. Antibodies and resins were crosslinked with 5 mM DSS for 1 hour at room temperature with continuous mixing, then the reaction was quenched with 1 M Tris HCl pH7.5 for 1 hour at room temperature with continuous mixing.

H-2Ld was immunoprecipitated from precleared proteins by continuous mixing with crosslinked resin/antibody at 4°C overnight, then the unbound protein extract was subsequently passed on the following crosslinked resin/antibody in order to immunoprecipitate H-2Kd/H-2Dd at 4°C overnight. The resins were washed and centrifuged for 2 times with 10 volumes of 400 mM NaCl, 20 mM Tris-HCl, 0.2% NP40 then with 15 volumes 20 mM Tris-HCl pH 8, 3 minutes each wash. Peptides were eluted from H-2 complexes with 8 washes in TFA 0.2%, 1 min each. Supernatants were passed through an Amicon Ultra 0.5mL 3k filter in order to separate H-2 molecules from the peptides.

Peptides in 0.2% TFA were dried by vacuum centrifugation, solubilized in 5% FA and purified by binding to disposable reversed-phase C18 stage tips. Samples were injected onto a quadrupole Orbitrap Q-exactive HF mass spectrometer (Thermo Scientific), each one in technical duplicate. Peptides separation was achieved on a linear gradient from 95% solvent A (2% ACN, 0.1% formic acid) to 55% solvent B (80% acetonitrile, 0.1% formic acid) over 120 minutes and from 55% to 100% solvent B in 3 minutes at a constant flow rate of 0.25 µl/min on UHPLC Easy-nLC 1000 (Thermo Scientific) where the LC system was connected to a 23-cm fused-silica emitter of 75 µm inner diameter (New Objective, Inc. Woburn, MA, USA), packed in-house with ReproSil-Pur C18-AQ 1.9 µm beads (Dr Maisch Gmbh, Ammerbuch, Germany) using a high-pressure bomb loader (Proxeon, Odense, Denmark). The mass spectrometer was operated in DDA mode as described previously (79): dynamic exclusion enabled (exclusion duration = 15 seconds), MS1 resolution = 70,000, MS1 automatic gain control target = 3×10^6, MS1 maximum fill time = 60 ms, MS2 resolution = 17,500, MS2 automatic gain control target = 1 x 10^5, MS^2 maximum fill time = 60 ms, and MS2 normalized collision energy = 25. For each cycle, one full MS1 scan range = 300-1650 m/z, was followed by 12 MS^2 scans using an isolation window of 2.0 m/z.

The MS data were analyzed using MaxQuant with 1% false discovery rate (FDR). Peptides were searched against the uniport-proteome_Mouse_010419 database or the customized reference databases that contained the sequences identified by RNAseq data. N-term acetylation and methionine oxidation were set as variable modifications. Enzyme specificity was set as unspecific when peptides were searched against the UniProt mouse database, while enzyme specificity was set as no enzyme when peptides were searched against customized reference databases and peptides FDR was set to 0.01.

Whole Genome Sequencing analysis

Genomic DNA (gDNA) was extracted from BALB/c tissue, *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cell lines using ReliaPrep gDNA tissue miniprep system (Promega). Starting from 500 ng of gDNA, Next Generation Sequencing (NGS) libraries were prepared in house by means of Nextera DNA Flex Library Prep kit (Illumina Inc., San Diego, CA, USA), according to the manufacturer's protocol. Quality of libraries was checked with High-Sensitivity DNA assay kit (Agilent Technologies, Santa Clara, CA), while DNA fragments' size distribution was assessed using the 2100 Bioanalyzer with a High-Sensitivity DNA assay kit (Agilent Technologies, Santa Clara, CA). Equal amounts of final DNA libraries were pooled and sequenced on NovaSeq 6000 (Illumina Inc., San Diego, CA, USA) as paired-end 150 bp reads at IntegraGen SA (Evry, France) and FastQ files were generated using bcl2fastq v2.17 software. Genomic analyses were performed using a bioinformatic pipeline previously described (74). On average, sequenced samples reached a median depth of 93x (Table 1). CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ mutational calling was performed subtracting BALB/c germline variants. Only genomic positions present with a minimum depth of 10x and supported by at least 9 mutated reads were examined. To annotate alterations at genomic level, a Browser Extensible Data (BED) file was built that included all genomic regions. Coding, intronic and UTR regions BED files were downloaded from the University of California Santa Cruz (UCSC) table browser (assembly: mm10; table: refFlat). Non-coding RNA (ncRNA) regions were extrapolated from the whole mm10 refFlat table filtering for $cdsEnd - cdsStart = 0$. Each of those specific region BED files was further processed with the *bedtools merge* command (80). To generate the BED file for the extragenic regions, the previously merged BED files were concatenated and subtracted from the whole mm10 chromosome annotation tracks. The size of each region was calculated using the *bedtools coverage -hist* command. The combination of coding, intronic and UTR merged tracks together with the extragenic regions BED file was employed for the mutational annotation. Normalized TMB was evaluated as the number of variants per megabase (Mb) considering those derived from each specific region. The analysis of edited mutations was performed calculating for each region the natural logarithm of normalized TMB ratios.

RNA Sequencing analysis

Total RNA was extracted from CT26 *Mlh1*$^{+/+}$ and CT26 *Mlh1*$^{-/-}$ cells using Maxwell® RSC miRNA Tissue Kit (AS1460, Promega), according to the manufacturer's protocol. The quantification of RNA was performed by DeNovix Ds-11 Spectrophotometer (Resnova) and Qubit 3.0 Fluorometer (Life Technologies). RNA integrity was evaluated with Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Kit. 500 ng of total RNA, with RNA integrity number (RIN) score between 8 and 10, was used for NGS Library using TruSeq Stranded mRNA Library Preparation Kit LP (48 samples) according to the manufacturer's protocol. The standard RNA fragmentation profile was used (94 °C for 8 mim). PCR-amplified RNAseq library quality was assessed using the Agilent DNA 1000 kit on the Agilent 2100 BioAnalyzer and quantified using Qubit 3.0 Fluorometer (Life Technologies). Libraries were diluted to 10 nM using Tris-HCl (10 mM pH 8.5) and then pooled together. The 7.5 pM diluted pool was run on MiSeq to evaluate library quality and balancing. Rebalanced pool was denatured according to the NextSeq system guide, and 1.3 pM were run on NextSeq500 using NextSeq 500/550 High Output v2.5 kit (150 cycles).

To calculate the coverage over-depth data, single-end FastQ files were processed as follows: files were aligned with MapSplice v2.2.0 (81) using mm10 assembly as reference genome. The generated alignment files were handled to translate genomic coordinates to transcriptomic ones and to filter out alignments carrying indels using the *sam-xlate* and *sam-filter* commands from UNC-Chapel Hill Bioinformatics Utilities. The final compressed Sequence Alignment/Map (BAM) files were inspected through the *bedtools genomecov* command using *-bga* and *-split* as parameters (80). The generated files were further analyzed using the *bedtools intersect* command (80) to count for every genomic region the number of bases covered for each minimum depth value. For each region, the count of annotated MAPs was normalized using the number of bases covered with at least 10x depth.

To calculate the peptide transcripts per million (TPM) the following formula was applied for each region: $TPM = \frac{P}{\sum(P)} \times 10^6$ where $P = \frac{number\ of\ MAPs\ supporting\ reads\ mapped\ to\ each\ region \times 10^3}{region\ length\ in\ base\ pair}$.

Generation of specific peptide database for mass-spectrometry data search

Each sequence contained in the FastQ files generated during the RNAseq experiment (Table 2) was subjected to a six-frame translation: the three possible reading frames in both directions of the strand. All the translated sequences were divided into KMERs of length 8-11 and then uniquely counted. The *Mlh1*$^{+/+}$ specific database was built including KMERs (peptides) that exhibited at least 10 counts at the time of injection and after excision from immunocompromised mouse (NOD-SCID), and that disappeared in tumor masses obtained from immunocompetent mouse (BALB/c). The *Mlh1*$^{-/-}$ custom database was assembled as follows: first peptides that showed at least 10 counts at the time injection and retained after excision from immunocompromised mouse were selected; then those peptides were compared to the sequences obtained in tumors growth in three immunocompetent mice. Peptide lost or strongly counter selected in at least one BALB/c tumor, measured as $(counts_{BALB/c} = 0\ or\ (counts_{pre-selected} - counts_{BALB/c} \geq 10\ and\ \frac{counts_{pre-selected}}{counts_{BALB/c}} \geq 10)\ )$, were further selected. Peptide sequences found in *Mlh1*$^{+/+}$ tumors excised from BALB/c and NOD-SCID were excluded from the *Mlh1*$^{-/-}$ specific database as well as sequences present in *Mlh1*$^{+/+}$ cells or strongly expanded in *Mlh1*$^{-/-}$ compared to the *Mlh1*$^{+/+}$ counterpart. The latter measure was calculated as follows:

$(counts_{CT26-Mlh1-} - counts_{CT26-Mlh1+} \geq 10 \; and \; \frac{counts_{CT26-Mlh1-}}{counts_{CT26-Mlh1+}} \geq 10)$.

MHC-I associated peptide annotation

Peptides identified by matching RNAseq database and the immune-peptidomic pipeline were further inspected to determine the genomic regions from which those peptides originated. For both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ peptide list, the original read name, the sequence, and frame of translation were retrieved. Next, the relative positions of the peptides inside the reads were calculated. Fasta files were generated and fed to *blat* (82) to retrieve genomic coordinates of the identified peptides. A score was calculated from the *blat* annotated files as follows: $(match + rep.match - mis - match - Q\_gap\_count - T\_gap\_count)$. For each read only the best score output was selected, and a BED file was generated with the determined genomic coordinates. The latter BED file was further examined through the *bedtools merge* command using *-d 5 -c 4 -o distinct,count* as parameters (80). Next, the resulting file was matched with the BED file that includes all the genomic regions previously described using the *bedtools intersect* command (80). Only uniquely mapped peptides were selected. In case peptide reads were aligned to regions that were not uniquely annotated, the following priorities were assigned to the genomic regions: 1) coding sequence; 2) 5'UTR; 3) 3'UTR; 4) intronic; 5) extragenic. Next, the annotated peptides were matched with the variant calling files to check the presence of SNVs and indels. Finally, all the peptides were examined combining the information from the canonical transcripts, generated from the UCSC refFlat table, to identify in-frame and out-of-frame peptides. The analysis of edited MAPs in *Mlh1*$^{-/-}$ tumor masses excised from BALB/c mice was performed calculating the log$_n$ fold change from pre-injection of RNA read counts + 1.

Results

Mutational and neoantigen landscape of MMRp and MMRd CRC cells

Intrigued by the efficacy of immunotherapy in MMRd tumor patients, we studied how MMR altered genes affect the genomic landscape of tumors and what is the peculiar contribution of each gene. To this end we genetically inactivated MMR genes such as *Mlh1, Msh2, Msh6* and *Pms2* - by using the CRISPR-Cas9 technology. To avoid off-target effects we selected two different clones generated with two different and independent guides for each gene. In parallel, two MMR *wild type* clones were selected as control. Each MMRp and MMRd clone was cultured *in vitro* for many days and underwent WES at 30, 90 and 150 days after the MMR inactivation. Finally, the mutational and neoantigenic characterization was generated using a bioinformatic pipeline previously described (74). Data analysis performed during the past two years revealed that the inactivation of *Mlh1, Msh2, Msh6* and *Pms2* in CT26 leads to the acquisition of several mutations – both SNVs and frameshift indels - during time as compared to the two MMRp clones (Figure 8 and 9).

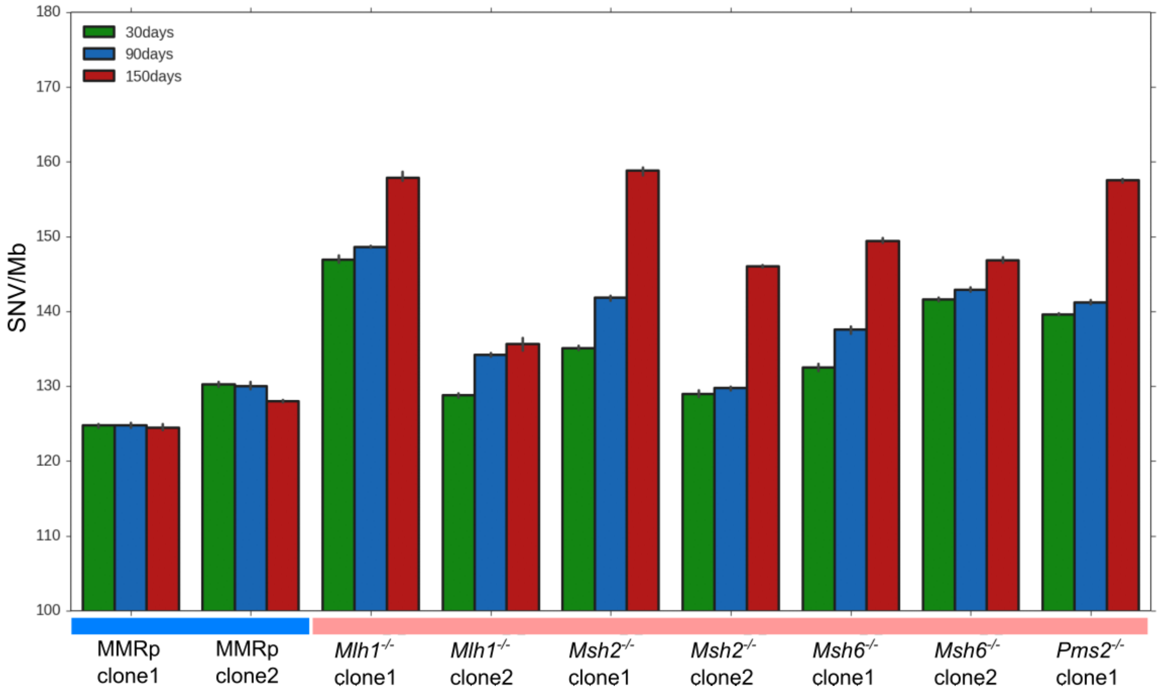

Figure 8 SNV landscape of MMRp and MMRd CRC cell models. Genomic analysis pipeline (see methods) was applied to MMRp (blue x-axis) and MMRd (red x-axis) CRC cells at 30 (green), 90 (blue) and 150 (red) days after MMR inactivation.

Notably, even if the absolute number of indels was lower than the number of SNVs, the relative increase of MMRd cells compared to MMRp cells was higher in the number of indels (Figure 9).

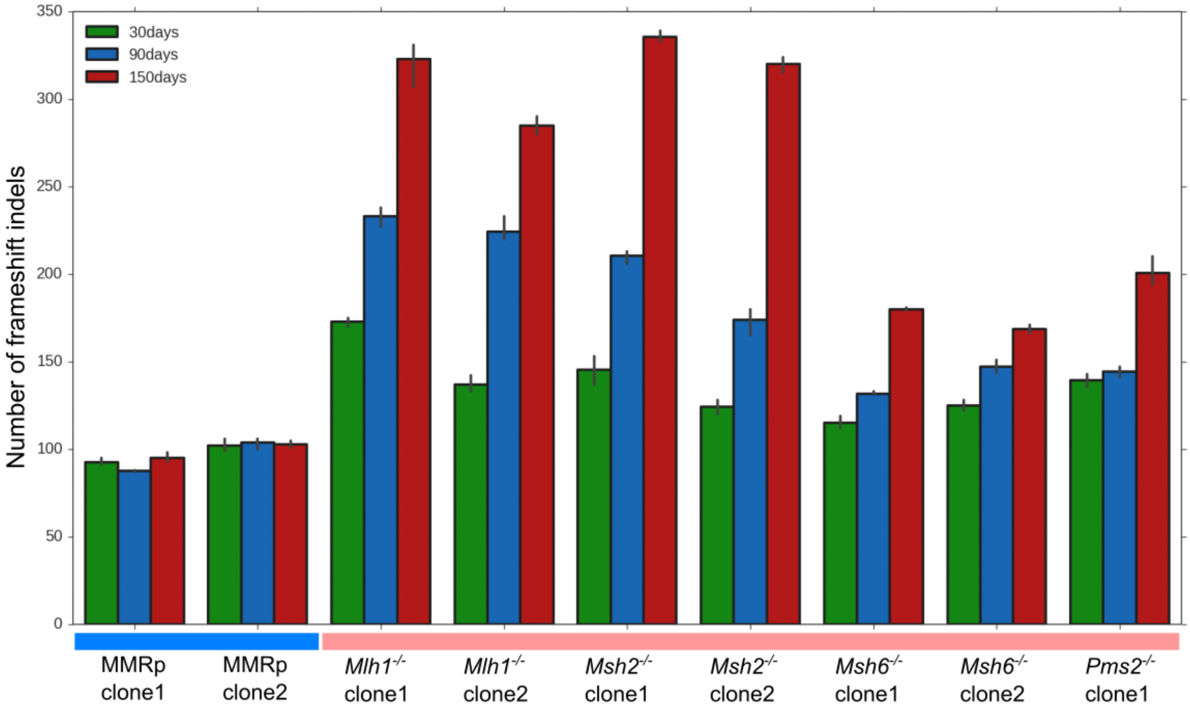

Figure 9 Frameshift indels landscape in MMRp and MMRd CRC cell models. Genomic analysis pipeline (see methods) was applied to MMRp (blue x-axis) and MMRd (red x-axis) CRC cells at 30 (green), 90 (blue) and 150 (red) days after MMR inactivation.

Moreover, we employed SNVs and indels information to perform the neoantigen prediction pipeline on the same dataset (4) and the results revealed that all MMRd acquired more putative neoepitopes over time compared to MMRp counterparts (Figure 10).

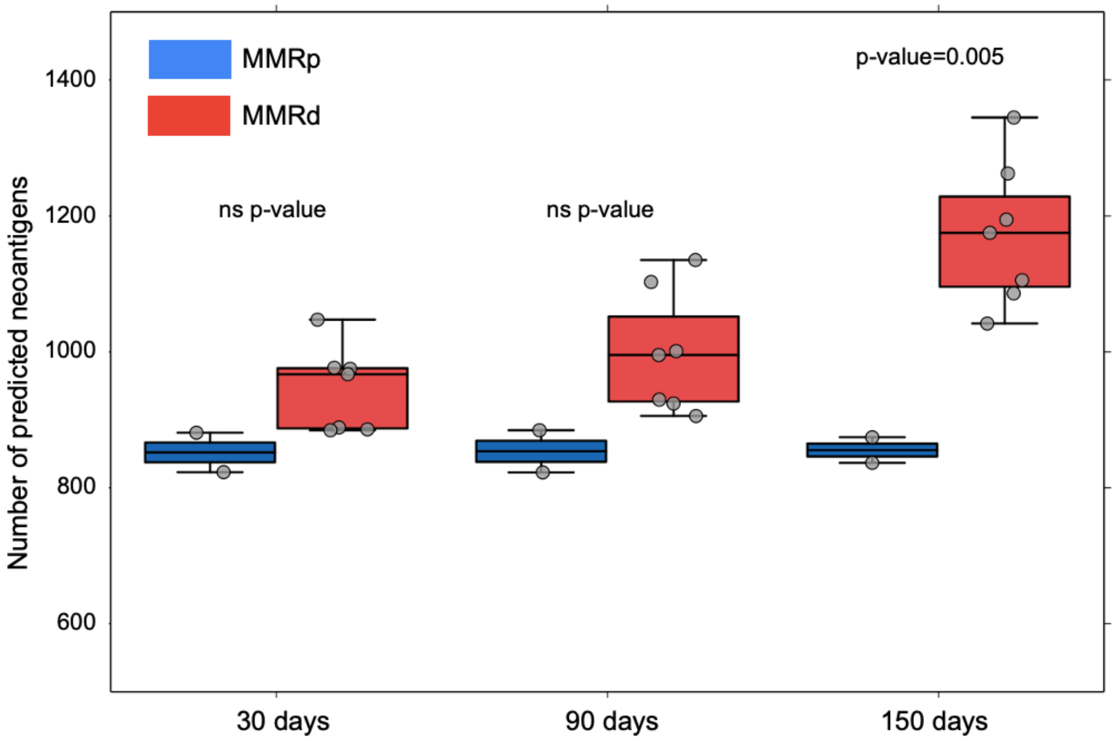

Figure 10 Number of predicted neoantigens acquired over time in MMRp and MMRd CRC cell models. Neoantigen prediction pipeline (see methods) was applied to MMRp (blu) and MMRd (red) CRC cells at 30, 90 and 150 days after MMR inactivation. Results were grouped according to MMR status and statistical test was performed at each time point (Independent samples T-test: ns= not significant).

Moreover, we calculated the relative frequency of SNV- and indel-derived neoantigens and the results highlighted that *Mlh1*$^{-/-}$ and *Msh2*$^{-/-}$ CRC cells showed the highest presence of predicted neoantigens generated from frameshift indel variants (Figure 11).

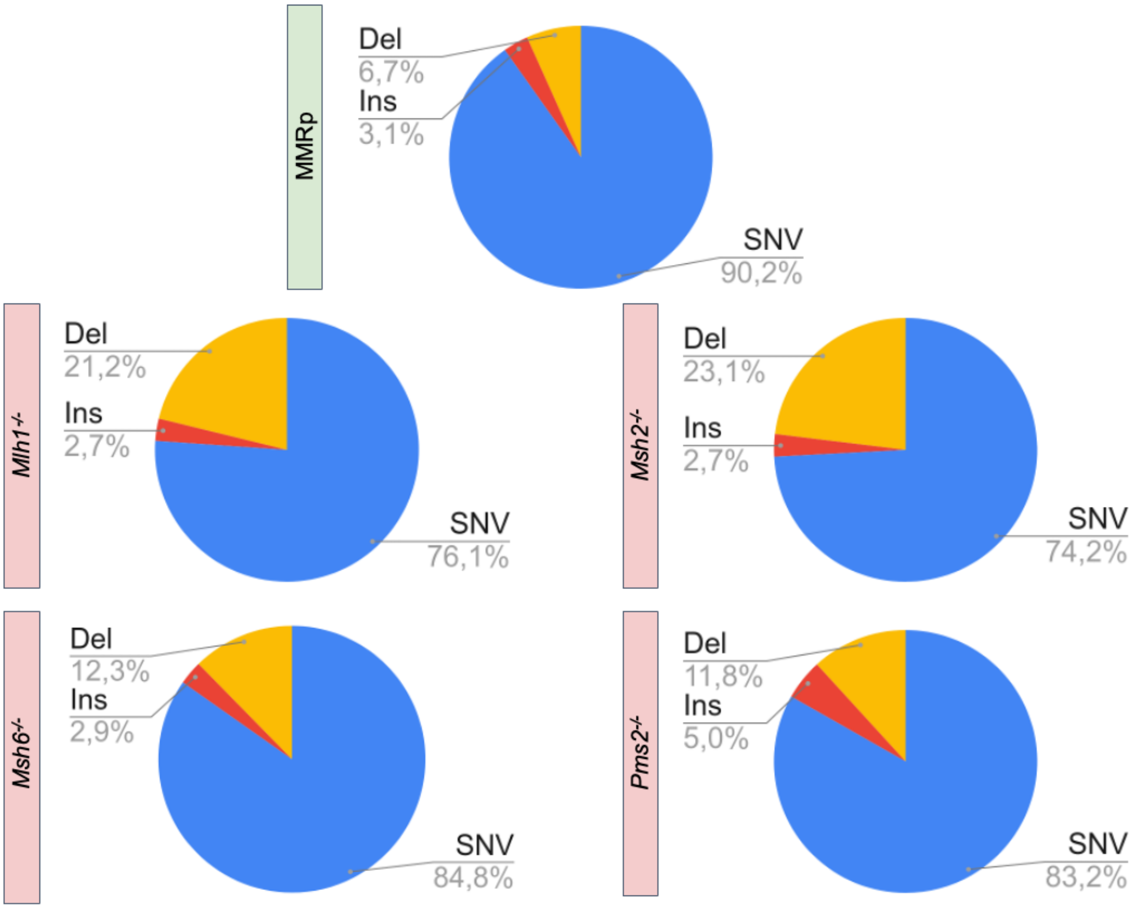

Figure 11 Relative frequency of neoantigen sources in MMRp and MMRd CRC cell model after 150 days from MMR inactivation. The percentage of alteration identity from which the neoantigens were predicted are shown for MMRp (green) and MMRd (red) CRC cells after 150 days from MMR inactivation.

Mutational signature profiles of MMRp and MMRd CRC cell models

Somatic mutations are a consequence of several endogenous or exogenous mutational processes such as exposure to ultraviolet light, tobacco carcinogens, treatment with alkylating agents and defect of DNA repair mechanisms. To investigate which are the mutational patterns acquired during tumor progression several Institutes collaborated into generate and maintain a database of mutational signatures (83). The most recent version of the mutational signature catalogue reports several units associated with DNA mismatch repair and microsatellite instability. Specifically, among them there are seven single base substitution (SBS) signatures: SBS6, SBS15, SBS20, SBS21, SBS26 and SBS44 (18). We reasoned that the noteworthy accumulation of genomic variants observed in MMRd CRC cells over time might be reflected in their mutational signatures. To examine that, we generated the mutational profiles of our set of CRC cells 150 days after MMR inactivation according to the computational strategy employed to reveal mutational signatures. MMRp cells showed a mutational signature mainly composed by C>T changes (Figure 12).

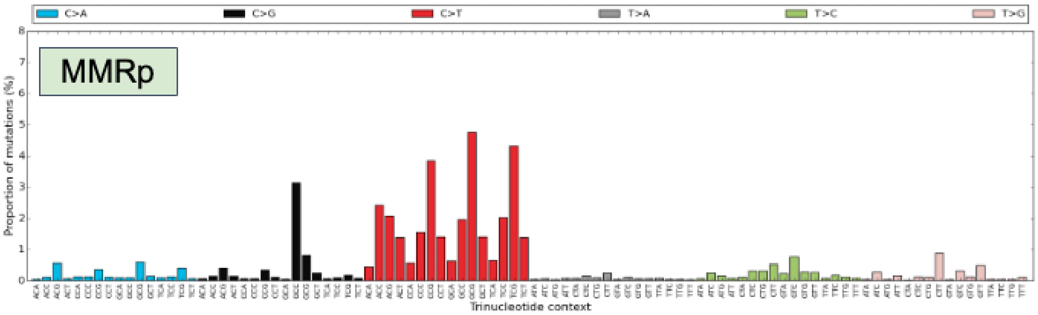

Figure 12 Mutational signature profile of MMRp cells. The MMRp SBS spectra consisting of 96 different contexts generated form the pyrimidines of the Watson-Crick base pairs - C>A, C>G, C>T, T>A, T>C, and T>G - and the bases immediately 5’ and 3’.

Notably, MMRd cells showed an increase of a specific trinucleotide context among the C>T changes, that is GCG> GTG, and a modest but sprinkled increase of T>C (Figure 13).

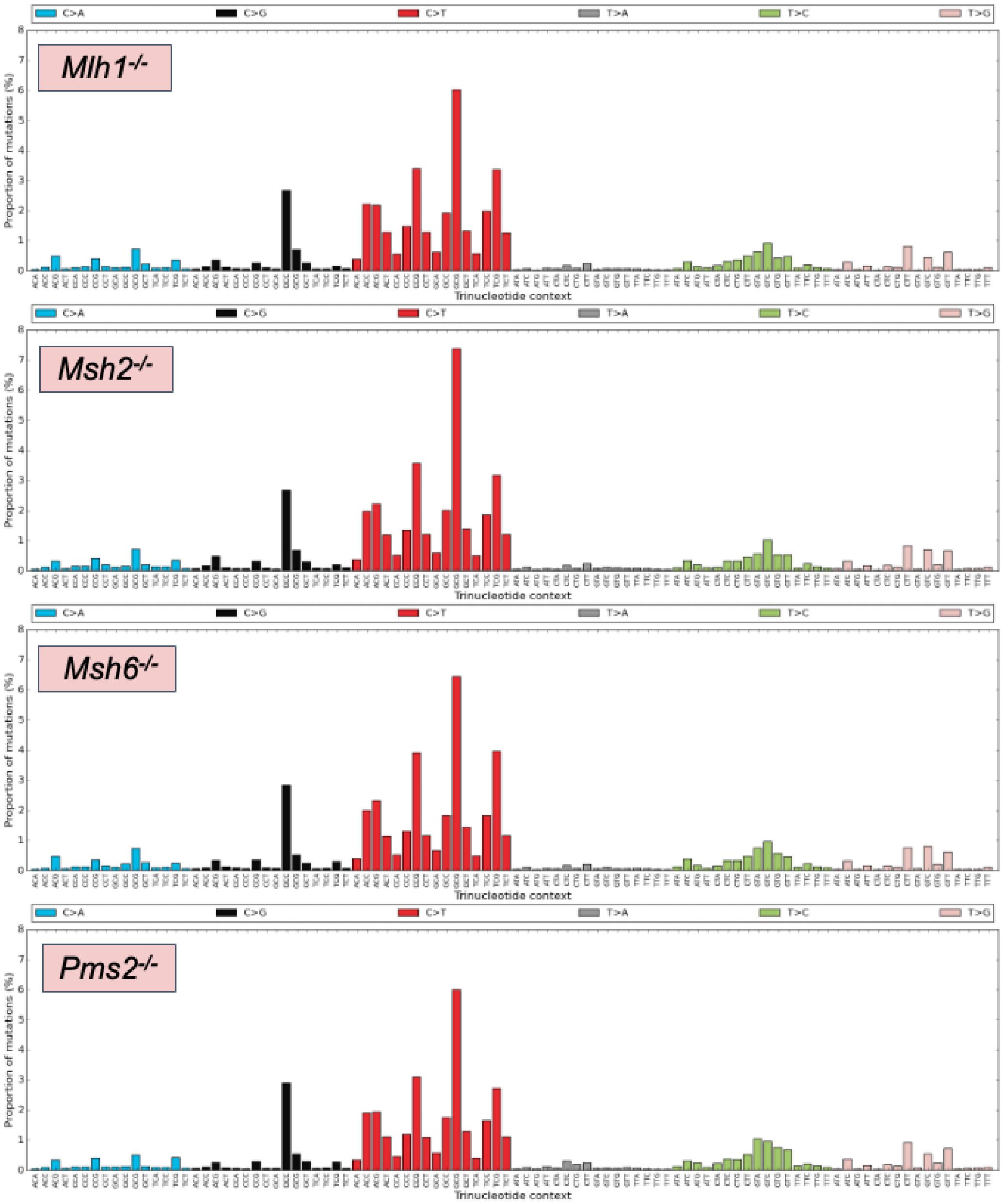

Figure 13 Mutational signature profile of MMRd cells. The MMRd SBS spectra consisting of 96 different contexts generated form the pyrimidines of the Watson-Crick base pairs - C>A, C>G, C>T, T>A, T>C, and T>G - and the bases immediately 5' and 3'.

Next, to confirm our hypothesis we exploited *signal,* an advanced computational tool, (78), to identify the specific mutational spectrum. Therefore, we decided to identify the mismatch repair associated mutational signature on our set of MMRp and MMRd CRC samples at 30, 90 and 150 after MMR inactivation. The results revealed that the number of mutations contributing to the generation of MMR specific mutational signature was stable over time (Figure 14). Conversely, MMRd CRC showed a higher number of mutations associated with MMR specific signature that increased over time (Figure 14).

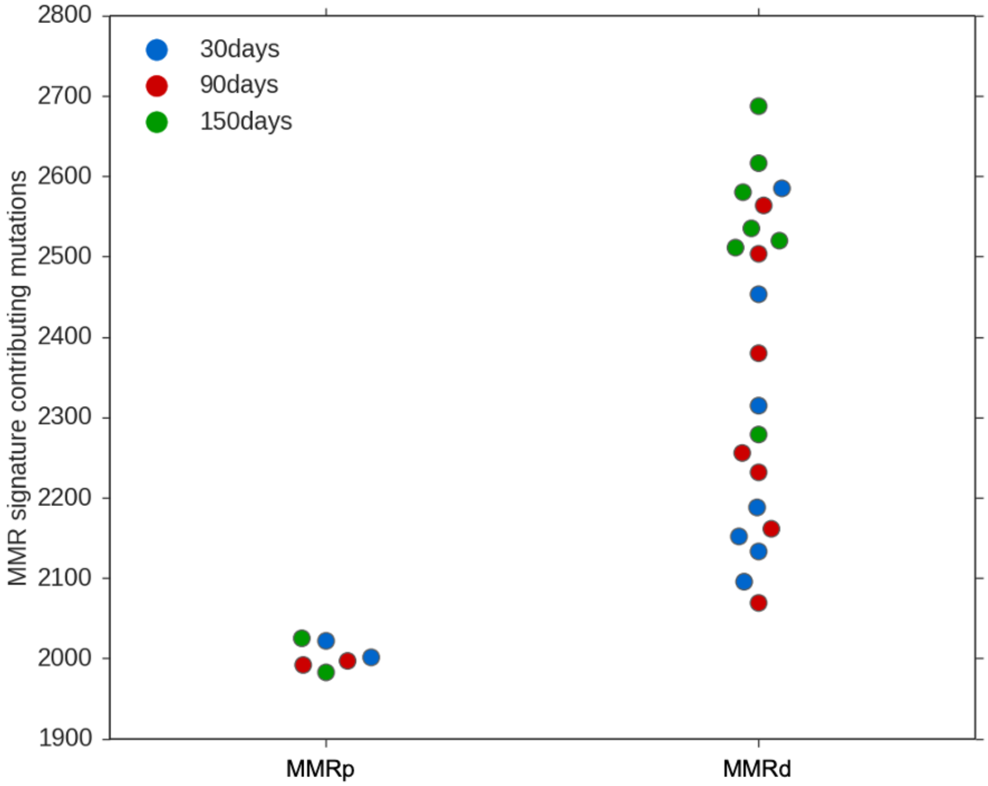

Figure 14 Prevalence of mutations that specifically contributes to MMR damage associated signature. Mutational signature profile was extracted using *signal* (see methods) and the number of alterations that contributed to the generation of MMR damage associated signature are reported for MMRp and MMRd CRC cells.

Characterization of edited alterations in MMR-proficient and -deficient CRC cells

To characterize the landscape of alterations edited by the immune system when the cancer cells were grown *in vivo*, we exploited previously described syngeneic mouse models (8, 77). We examined the impact of the immune system on cancer cells by injecting MMR-proficient and -deficient cells into immunocompromised (NOD-SCID) and immunocompetent mice (BALB/c) (Figure 15A).

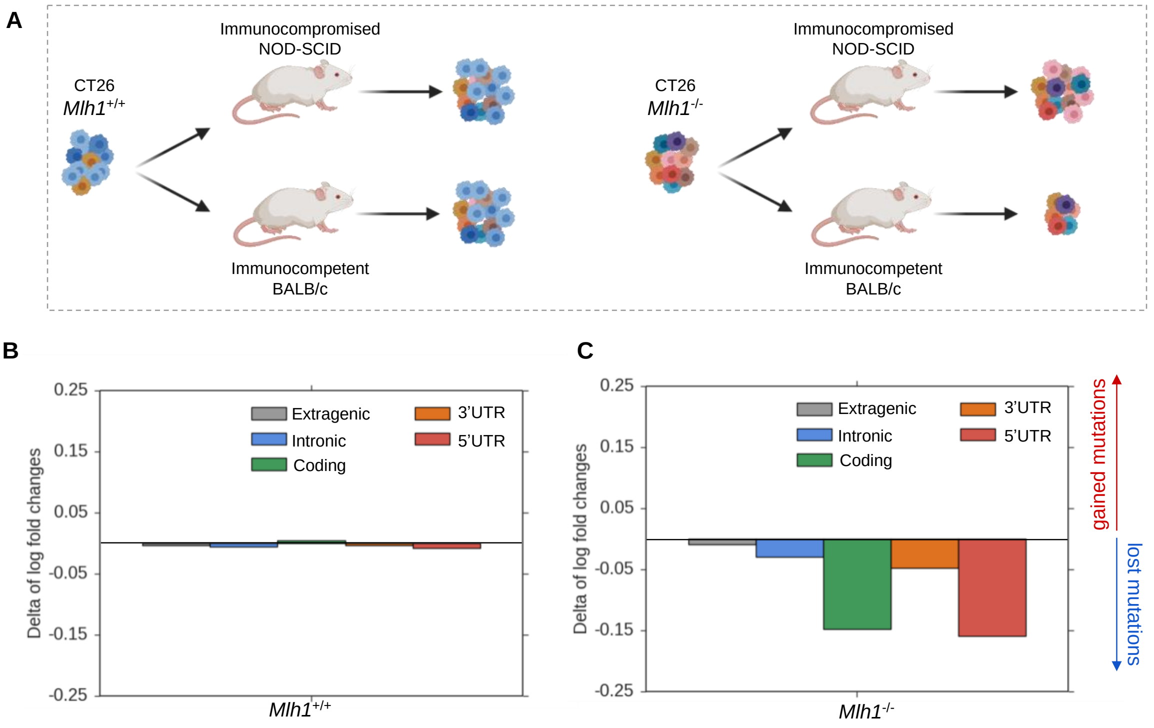

Figure 15 Analysis of edited mutations in MMR-proficient and -deficient CT26 after injection in immunocompromised and -competent mice. **(A)** Experimental workflow employed for the analysis of edited mutations in WGS data of CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples. Briefly, each CT26 clone was inoculated into NOD-SCID (immunocompromised) and BALB/c (immunocompetent) mice. CT26 MMR-proficient and -deficient tumors underwent WGS at the time of injection and after excision from the mice. Delta between log fold changes evaluated after injection in immunocompromised and -competent mice in CT26 Mlh1$^{+/+}$ **(B)** and CT26 Mlh1$^{-/-}$ **(C)**. Log fold changes analysis of gained and lost alterations were calculated from CT26 Mlh1$^{+/+}$ and Mlh1$^{-/-}$ pre-injection data respectively. The alterations were grouped in regions and normalized per Mb before log fold change calculation.

Tumor cells were subjected to high depth WGS at the day of mouse implantation and at the time of excision, i.e., after 15 days of growth in mice (Table 1).

Sample	**Reads**	**Mapped:**	**Median depth:**
BALB/c	1078015820	99.56%	105
CT26 *Mlh1*$^{+/+}$	958072282	99.61%	91
CT26 *Mlh1*$^{+/+}$ post BALB/c M1	877846209	99,66%	83
CT26 *Mlh1*$^{+/+}$ post BALB/c M2	856369151	99,63%	80
CT26 *Mlh1*$^{+/+}$ post BALB/c M3	1088138369	99.75%	103
CT26 *Mlh1*$^{+/+}$post NOD-SCID M1	915507037	99.61%	85
CT26 *Mlh1*$^{+/+}$post NOD-SCID M2	964605317	99.62%	92

CT26 *Mlh1*$^{+/+}$post NOD-SCID M3	888087399	99.58%	82
CT26 *Mlh1*$^{-/-}$	1083850382	99.71%	104
CT26 *Mlh1*$^{-/-}$ post BALB/c M2	996468559	99.59%	90
CT26 *Mlh1*$^{-/-}$ post BALB/c M6	1038671151	99.52%	94
CT26 *Mlh1*$^{-/-}$ post BALB/c M7	1089980019	99.63%	106
CT26 *Mlh1*$^{-/-}$ post NOD-SCID M5	1022739553	99.63%	100
CT26 *Mlh1*$^{-/-}$ post NOD-SCID M7	854313322	99.59%	81

Table 1 List of WGS analyses performed in CT26 samples.

To assess whether and how the genomic profile of *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ tumor cells evolved in the presence of a competent or compromised immune system, we first evaluated the mutational landscape of each sample. More precisely, we calculated the number of alterations per Mb that emerged in each distinct genomic region (Figure 16A and 16B).

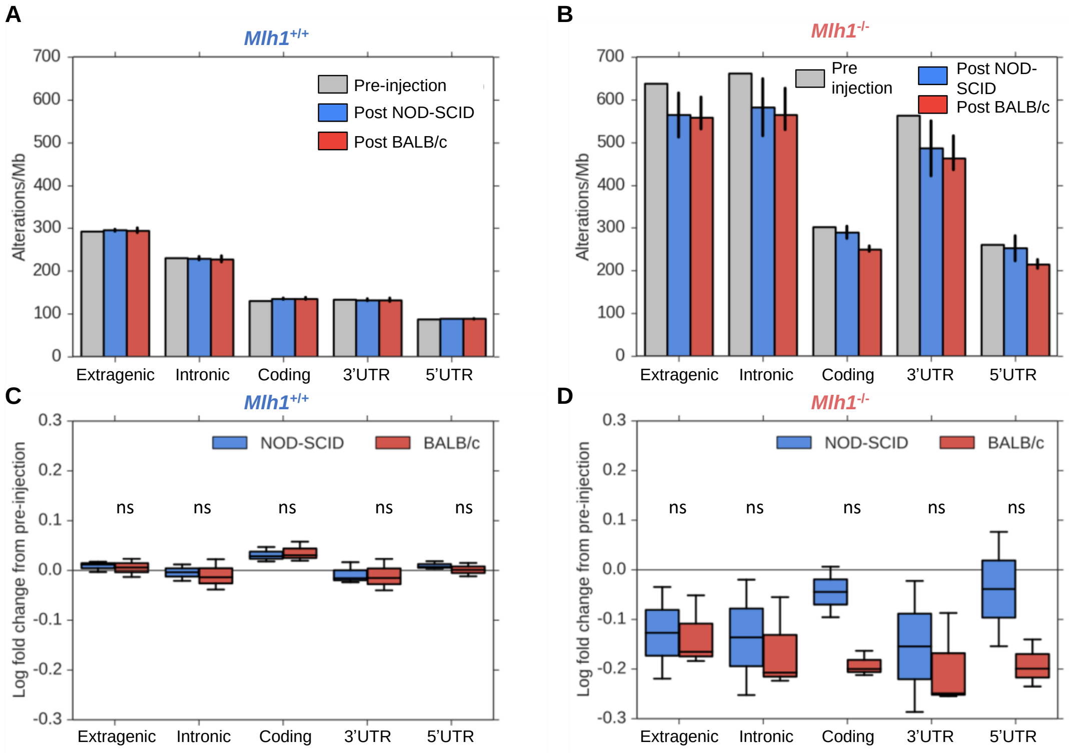

Figure 16 Characterization of alterations in CT26 samples before and after injection in immunocompromised and -competent mice. Number of alterations per Mb calculated in every genomic region of CT26 *Mlh1*$^{+/+}$ **(A)** and CT26 *Mlh1*$^{-/-}$ **(B)** pre- and post-growth in immunocompromised and -competent mice calculated by WGS. The alterations were grouped in regions and normalized per Mb. Log fold change analysis of gained and lost alterations from pre-injection evaluated after tumor growth

of CT26 *Mlh1*$^{+/+}$ **(C)** and CT26 *Mlh1*$^{-/-}$ **(D)** in immunocompromised and -competent mice. The alterations were grouped in regions and normalized per Mb before log fold change calculation. (Mann-Whitney U test: ns non-significant).

We found no differences in the mutational spectrum of MMR-proficient cells pre- and post-injection in both immunocompetent and -compromised mouse models (Figure 16A). Overall, these results showed no evidence of immune edited mutations in MMRp tumors grown in immunocompetent mice. Conversely, a considerable increase in alterations in all genomic regions of *Mlh1*$^{-/-}$ cells were observed, particularly in the non-coding areas (Figure 16B). Moreover, a considerable decrease in mutations per Mb was observed in MMR-deficient cells grown in immunocompetent mice compared to cells before the injection and grown in immunocompromised mice. This result prompted us to examine the contribution of the immune system against antigenic mutations; to this end, we calculated the mutational differences of MMR-proficient and -deficient cancer cells that grew in immunocompetent and -compromised mice. Specifically, we evaluated the log fold change from preimplantation cells of gain and lost mutations after tumors growth *in vivo* (Figure 16C and 16D). No differences were observed in gain and lost mutations in *Mlh1*$^{+/+}$ clones after injection in immunocompromised and -competent mice (Figure 15B). On the contrary, a marked shrinkage (log fold change) was evident in 5'UTR and coding regions of the CT26 *Mlh1*$^{-/-}$ genome (Figure 15C); overall these data suggest that sequences generated in those regions were supposed to be removed by the activity of the immune system.

Identification of MHC class I associated peptides

We studied the impact of the immune system on the mutational profile in the coding and non-coding regions at the protein level. We reasoned that non-coding regions affected by immune-editing in *Mlh1*$^{-/-}$ cells *s*hould have been transcribed, translated, and further processed to be presented as a (neo)antigens on the cell surface, allowing the generation of (neo)peptides from genomic non-coding areas. We hypothesized that the non-coding regions counter-selected by the immune system encompass unconventional MAPs induced by inactivation of the MMR pathway. To test this hypothesis, we developed an extensive neoantigen identification pipeline integrating whole genome and RNA sequencing and an immune-peptidomic investigation through mass-spectrometry analysis (Figure 17).

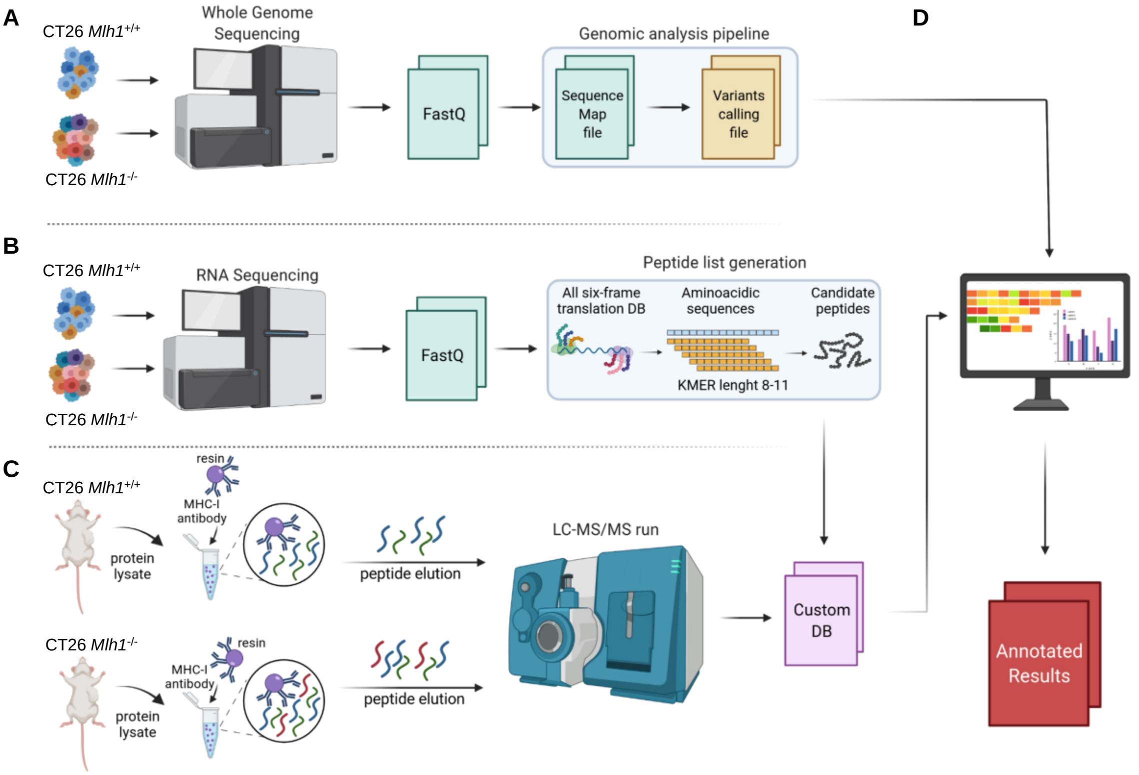

Figure 17 Pipeline design for MAP identification. **(A)** WGS data were generated from CT26 *Mlh1*$^{+/+}$ and Mlh1$^{-/-}$ samples and analyzed using IDEA pipeline (see methods) (74)in order to produce the alignment and variant calling files. **(B)** RNAseq data were further generated from CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples. FastQ files were handled to produce the list of all putative peptides present in the transcriptome of each sample. In brief, every transcript sequence in the FastQ files underwent all-six frame translation; then the lists of 8-11 amino-acid long peptides were generated using the KMER approach; finally, the peptide lists were compared to select only peptides edited in tumors excised from immunocompetent mice (see methods). **(C)** CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ tumor masses were explanted from NOD-SCID mice (n=6 per group) and protein extraction was performed. MHC-I was isolated from whole protein lysates through H-2d antibodies conjugated to resin, then peptides were eluted from MHC-I and injected in a mass spectrometer. The MS data were then analyzed using MaxQuant. Peptides were searched against the customized DB made of edited peptides generated by RNAseq data. **(D)** Sequence results obtained from the immune-peptidomic pipeline were ultimately matched with WGS data to retrieve information about the genomic sources of edited peptides (see methods).

As mentioned before, at first, we performed WGS on *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ before- and after growth in BALB/c and NOD-SCID mice (Figure 15A). Through genomic analysis of these samples, we generated the variant calling files. Then, RNA extracted from the same samples was also sequenced (Table 2).

Sample	**Mates**	**Mapped:**
CT26 *Mlh1*$^{+/+}$	62164033	98.73%
CT26 *Mlh1*$^{+/+}$ post BALB/c M3	61712173	98.20%
CT26 *Mlh1*$^{+/+}$post NOD-SCID M2	59494801	98.72%
CT26 *Mlh1*$^{-/-}$	53045412	98.75%
CT26 *Mlh1*$^{-/-}$ post BALB/c M2	57481236	98.92%
CT26 *Mlh1*$^{-/-}$ post BALB/c M6	59195273	98.82%
CT26 *Mlh1*$^{-/-}$ post BALB/c M7	58457417	98.80%
CT26 *Mlh1*$^{-/-}$ post NOD-SCID M5	60813574	98.88%

Table 2 List of RNAseq performed in CT26 samples.

However, the RNAseq FastQ files were not aligned to the reference transcriptome (in contrast to procedures previously performed for the genomic pipeline); instead, the raw data were used to create two different databases containing all the peptides that could originate from the transcripts of both the MMR-proficient and -deficient cancer cells (Figure 17B). Ultimately, we applied the immune-peptidomic pipeline to unveil the antigenic profile presented by the MHC class I on the surface of *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells (Figure 17C). Briefly, we performed MHC-I immunoprecipitation on protein lysates of both MMR-proficient and -deficient tumors grown in immunocompromised animals. Next, peptides eluted from the MHC-I molecules, were analyzed by LC-MS/MS, and searched against customized reference databases that contained all putative peptide sequences selected by RNAseq analysis. The specificity of each peptide was verified by a cross-check of the *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ databases on both samples: *Mlh1*$^{+/+}$ eluted peptides against the *Mlh1*$^{-/-}$ specific database and vice versa. Through this approach we selected only *Mlh1*$^{-/-}$ exclusive peptides (*Mlh1*$^{-/-}$ specific database), whilst all the peptides eluted from both MMR-proficient and MMR-deficient samples were included in the *Mlh1*$^{+/+}$ database. Finally, we merged the results from the mass-spectrometry analysis and the WGS analysis to characterize the mutational status and annotate the genomic origin of MAPs (Figure 17D).

Identification of edited peptides in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ tumor cells

To generate a peptide database for mass-spectrometry analysis we exploited two mouse models, with severely compromised (NOD-SCID) or proficient (BALB/c) immune systems. This strategy allowed us to select only peptides putatively edited by a functional immune system. The RNAseq analysis of tumor cells, from which we inferred the peptide sequences, revealed more than 2469 million possible amino acid sequences from *Mlh1*$^{+/+}$ transcripts (Figure 18A).

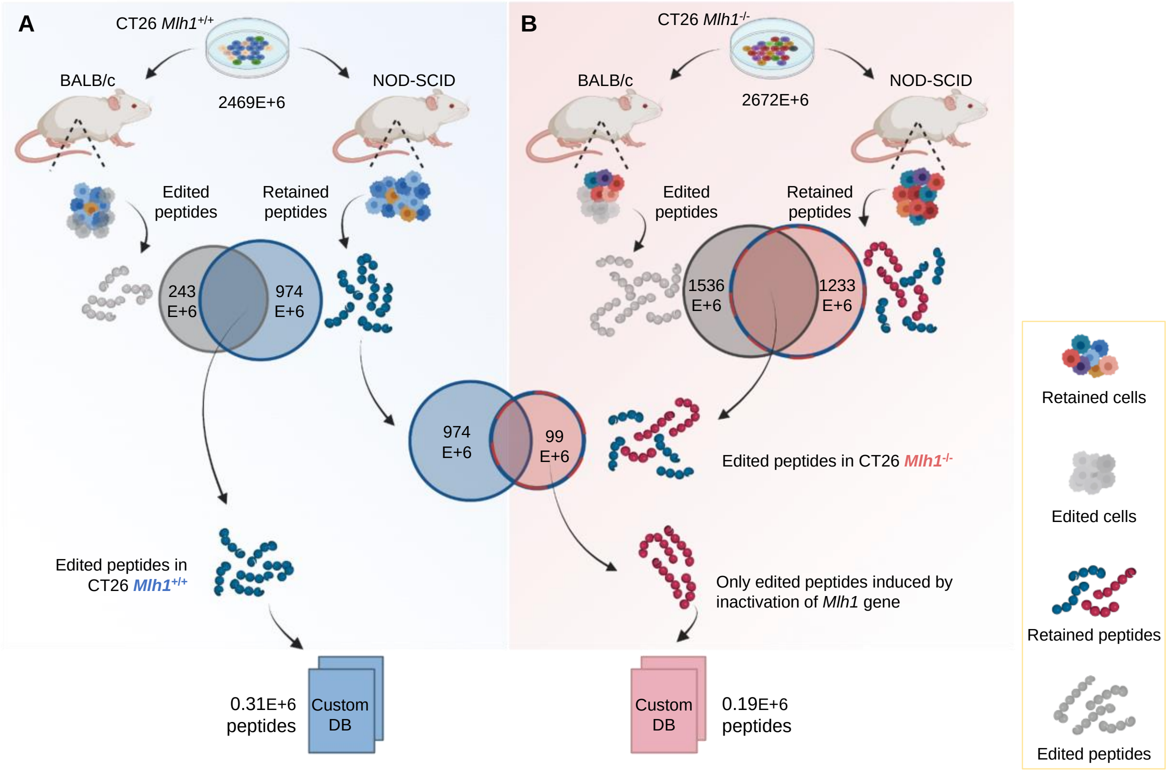

Figure 18 Identification of edited MAPs in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ tumor cells. **(A)** The peptide list generated from RNAseq analysis of CT26 *Mlh1*$^{+/+}$ cells grown *in vitro* was compared to the corresponding lists obtained after tumor growth in mice. Thus, peptides lost after injection in immunocompetent BALB/c mice and retrieved after inoculation in immunocompromised NOD-SCID mice were selected. The overlap of these two peptide datasets generated the database of CT26 *Mlh1*$^{+/+}$ edited peptides. **(B)** Peptide lists generated from RNAseq analysis in Mlh1$^{-/-}$ samples before and after in vivo growth were compared. This allowed the identification of peptides lost after injection in immunocompetent BALB/c mice but maintained in immunocompromised NOD-SCID mice. The overlap of these two datasets generated a list of peptides from which specific CT26 *Mlh1*$^{+/+}$ sequences were removed. The latter list created the edited peptides database specific to CT26 *Mlh1*$^{-/-}$.

To specifically gather the peptides that trigger a proficient immune activation, we selected the translated sequences that were counter selected after cell inoculation into immunocompetent animals and crossed these results with the sequence list retained after *Mlh1*$^{+/+}$ cells injection into immunocompromised mice. The combined results generated a list of 305506 peptides from which a custom database for *Mlh1*$^{+/+}$ cells was built. The same workflow was used for *Mlh1*$^{-/-}$ cells leading to the identification of 99 million sequences that were immune edited in the BALB/c mouse (Figure 18B). We excluded peptides also present in *Mlh1*$^{+/+}$ cells from this list, since our aim was to identify sequences induced by the inactivation of MMR machinery which were edited by the immune system. A total of 193312 sequences were identified in the *Mlh1*$^{-/-}$ custom database (Figure 18B).

Before applying the immune-peptidomic pipeline (Figure 17C), using the above-described custom databases, we verified cell surface levels of MHC class I in both MMR-proficient and -deficient cell models (Figure 19).

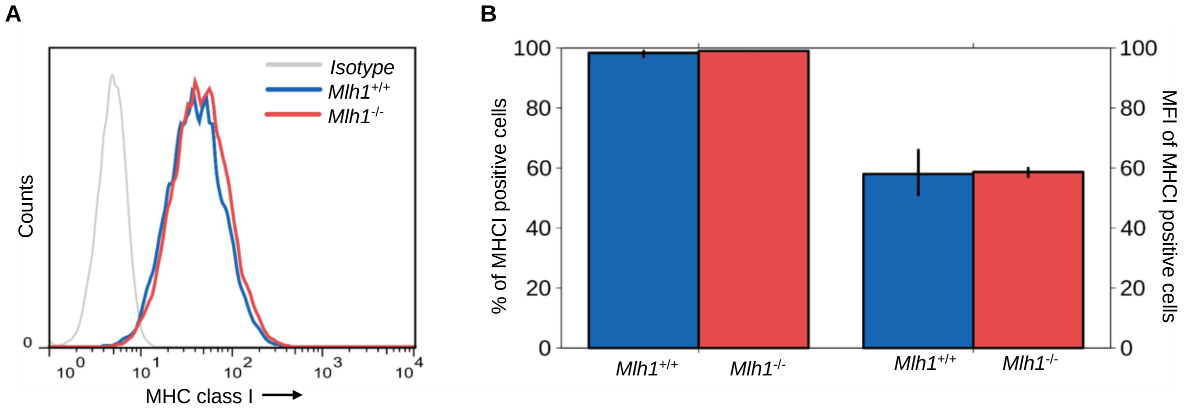

Figure 19 MHC class I levels in CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$. **(A)** MHC class I surface expression in CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ measured by FACS. The gray line corresponds to staining with an isotype control antibody. **(B)** The percentage of MHC class I positive cells and the mean of fluorescence intensity were depicted for three biological replicates.

The overall approach identified 417 peptides specifically exposed on *Mlh1*$^{+/+}$ surface, whilst 775 peptides were found to be specific of *Mlh1*$^{-/-}$ tumors (Table 3).

Sample	***Mlh1***$^{+/+}$ **or** ***Mlh1***$^{-/-}$ **custom database**	**Uniprot mouse database**
CT26 *Mlh1*$^{+/+}$ post NOD SCID	417	234
CT26 *Mlh1*$^{-/-}$ post NOD SCID	775	362

Table 3 List of LC-MS/MS run results in CT26 samples.

MAP classification in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ murine CRC cell line

To identify the genomic regions from which MAPs originated, we exploited WGS data. First, we investigated the resulting peptides among the translated DNA sequences generated in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells. Then, we ran a refining alignment of peptide derived from the specific reads, and we annotated them on the mouse genome. Next each genomic region was assigned to *coding, 3'UTR, 5'UTR, intronic or extragenic* labels. Moreover, according to the open reading frame of the sequences and the canonical isoforms annotated in the mouse transcriptome each peptide was further annotated as *in-frame* and *out-of-frame*. In total, we were able to confidently annotate 396 (95%) and 665 (86%) MAPs in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells respectively. Interestingly, our results showed that most MAPs derived from non-coding regions (Figure 20A).

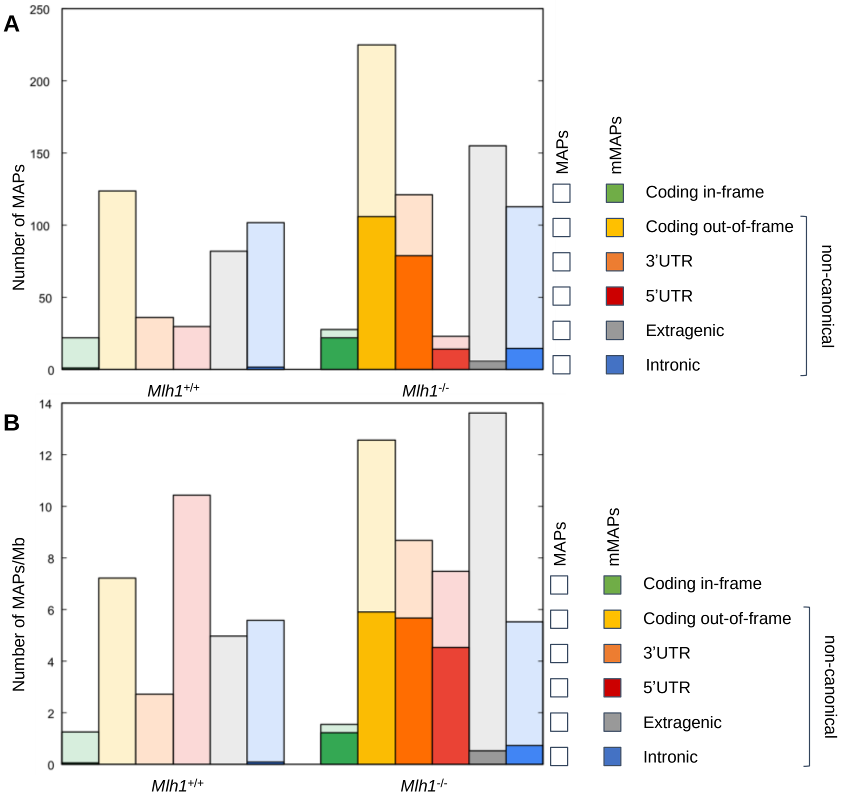

Figure 20 MMR-proficient and -deficient CT26 cells showed a high number of edited non-canonical MAPs. **(A)** The number of MAPs annotated at genomic level in CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples are reported in light colors. mMAPs are highlighted in solid colors. **(B)** The number of annotated MAPs was normalized (per Mb) in CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples and are reported in light colors. mMAPs are highlighted in solid colors.

More specifically, the majority of them were classified as non-canonical, since many MAPs, albeit originated from coding regions, showed *out-of-frame* translations in both MMR-proficient and -deficient cells. We took advantage of the variant calling files obtained from the genomic analysis pipeline to study which type of mutations (SNVs or indels) affected the MAPs. Notably, mMAPs were most abundant in *Mlh1*$^{-/-}$ cells and were mainly located in coding and UTR regions (Figure 20A). On the contrary, *Mlh1*$^{+/+}$ cells provided only a few mMAPs.

We considered that the polyA capture technique of RNA molecules for the subsequent RNAseq analysis could have perturbed the prevalence of MAPs in specific regions since they were better represented in the transcriptome. For this reason, we calculated the covered base pairs in each sample (Figure 21) and then re-evaluated the peptide classification after normalizing the data for this parameter. This analysis showed a higher prevalence of 5'UTR-derived MAPs per Mb in both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ models, while the overall trend of all other regions did not change (Figure 20B).

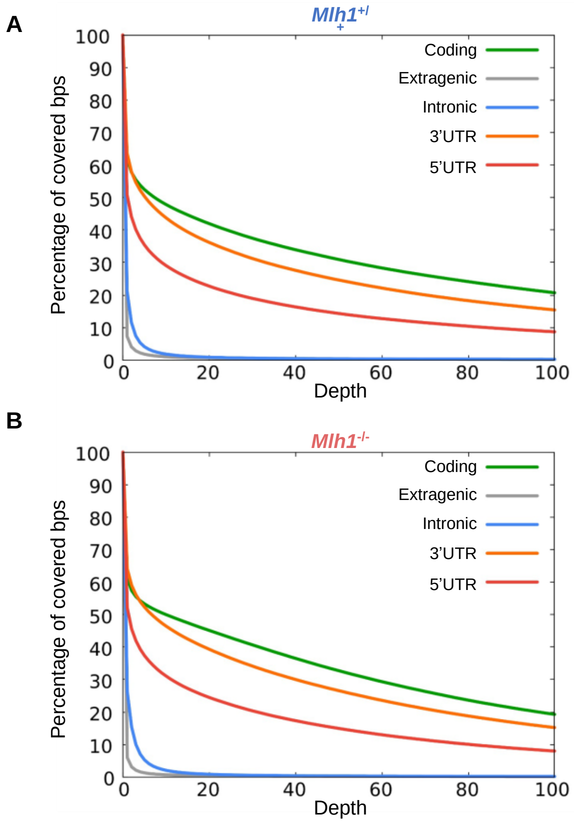

Figure 21 Coverage over depth analysis in all genomic regions of CT26 generated from RNA sequencing. The percentage of covered bases at single depth value resolution was calculated for each genomic region of *Mlh1*$^{+/+}$ **(A)** and *Mlh1*$^{-/-}$ RNAseq data **(B)**.

To assess the reliability of our workflow we decided to apply the immune-peptidomic pipeline against the UniProt mouse database (Figure 22A). We identified 171 mouse MAPs in common to both MMR-proficient and -deficient cells, while 63 and 191 exclusive MAPs were found in the *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ clones respectively (Table 3, Figure 22B). As expected, almost all the sequences present in the mouse canonical database were classified as coding.

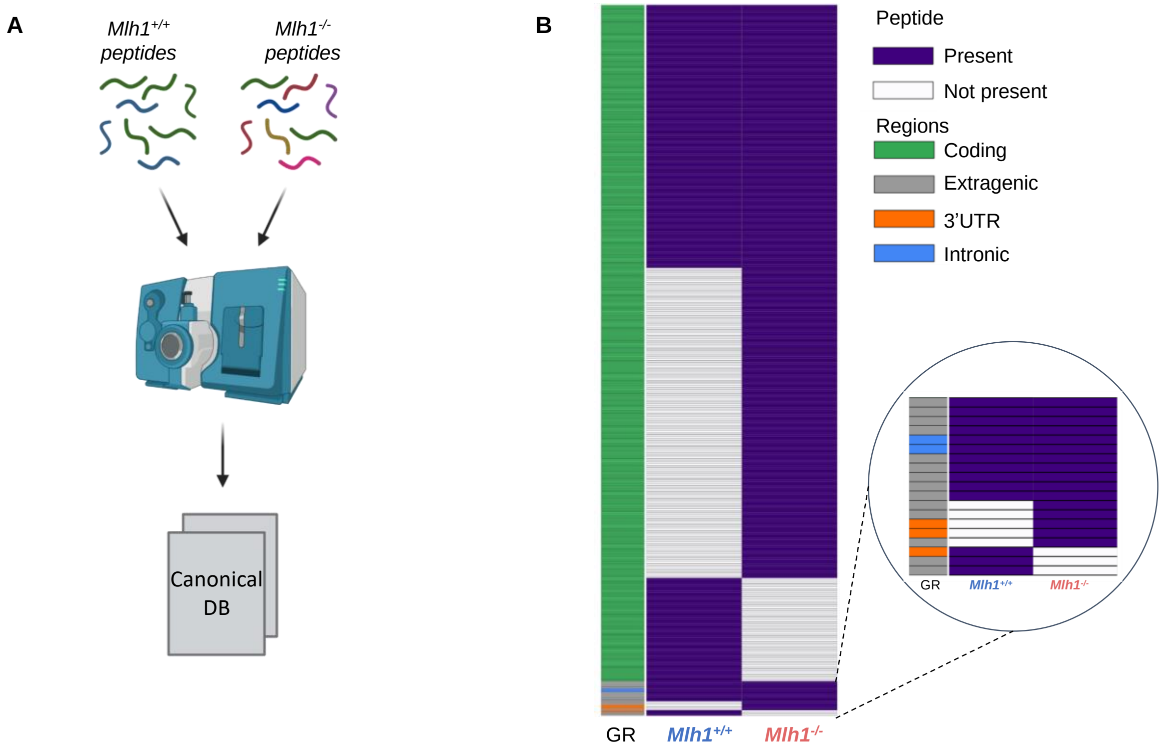

Figure 22 Characterization of the canonical peptide database. **(A)** Streamlined workflow of the immune-peptidomic pipeline in which MS spectra are matched to the UniProt mouse database. **(B)** Heatmap showing the peptide calls in CT26 *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples searching throughout the UniProt mouse database. The first heatmap column displays the genomic region (GR) annotation.

To further corroborate our findings, we determined the change in the expression levels of each edited MAP in *Mlh1*$^{-/-}$ tumors grown in immunocompetent mice, considering the biological variability of expression across different animals. To this end, we evaluated the quantity of RNA sequences supporting the peptide calls in CT26 *Mlh1*$^{-/-}$ before- and after-growth in immunocompetent animals and then we calculated the log fold change (Figure 23A). Notably, mMAPs exhibited a lower expression in *Mlh1*$^{-/-}$ cells grown in mice as compared to *wild type* MAPs, suggesting that those sequences were potently and efficiently targeted by the immune system of the host. Indeed, fold change analysis revealed a statistically significant reduction of mMAP transcripts (grouped in coding and non-coding) compared to *wild type* MAPs in *Mlh1*$^{-/-}$ cells grown in immunocompetent mice (Figure 23B).

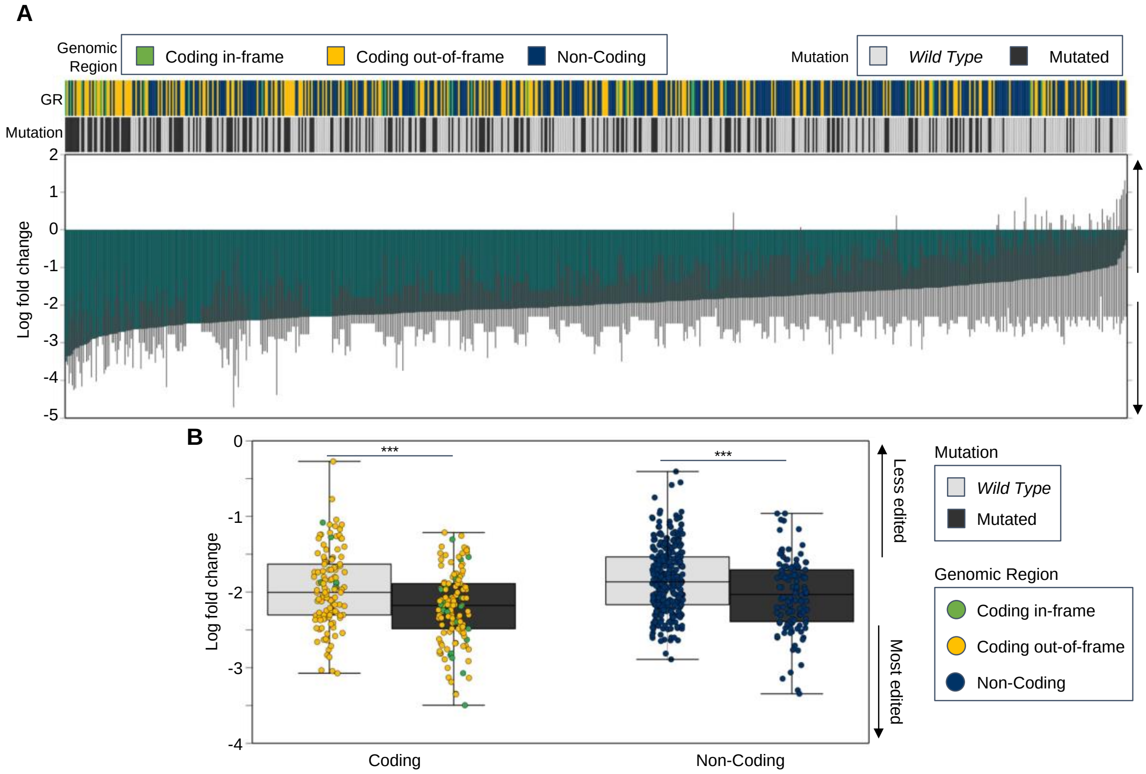

Figure 23 Immune editing of mutated MAPs in CT26 *Mlh1*$^{-/-}$. **(A)** Log fold change analysis performed between transcript values of CT26 *Mlh1*$^{-/-}$ peptides at the time of injection over those found after tumor excision from immunocompetent mice (dark green bars). The values are sorted from the lowest, i.e., the most edited MAPs, to the largest one, that is the least edited. In gray, the standard deviation among the three mice measurement is reported. **(B)** Log fold changes from pre-injection values were grouped according to the peptide mutational status (Independent samples T-test: *** p-value < 0.0005).

Discussion

Although molecular defects in the MMR machinery may be considered an escape strategy that leads cancer to a rapid evolution and uncontrolled dissemination, considerable evidence has highlighted how this is a double-edged sword for tumor cells (84). We previously showed that MMRd tumors trigger a remarkable immune response owing to their high neoantigen burden (8). We reported that a higher CD8$^+$ T-cell infiltration was present in the tumor microenvironment alongside a high number of distinct TCR rearrangements in blood of tumor-bearing mice (8).

In this work, we used CT26 mouse cell model to understand whether and to what extent alterations of DNA repair products modulate neoantigen profiles over time and how these can contribute to the immunogenic properties of MMRd CRCs. We exploited the CRISPR-Cas9 technology to genetically inactivate four key components of MMR pathway - *Mlh1, Msh2, Msh6* and *Pms2* – and consequently generate four MMRd CRC mouse cell models. Together with two MMRp CRC cells, they were passaged *in vitro* several times and WES data were generated. CRC cells carrying MMR defects accumulated a great number of mutations which led generating novel predicted peptides. Among the others, *Mlh1*$^{-/-}$ and *Msh2*$^{-/-}$ cells presented the highest prevalence of neoantigens after 150 days from the MMR inactivation. Moreover, those cells also showed the highest incidence of indel-derived neoantigens.

It has also been reported that somatic mutations accumulated throughout cell activity may leave a fingerprint on the DNA. Indeed, malfunction of the DNA repair system could affect not only the quantity but also the quality of mutations. Therefore, we deeply examined the mutational pattern acquired in MMRp and MMRd cells and our results revealed that MMRd induced mutations clearly defined a specific signature profile associated with DNA repair damages.

Our analysis was performed in line to the conventional strategy used in research and clinical practice by inspecting the coding DNA that defines the landscape of tumor neoantigens. However, it is currently unknown whether and to what extent the non-canonical neoantigen landscape, sometimes referred to as the “dark” side of the genome (i.e., the non-coding part), plays a role in the immunogenic features of MMRd tumors. Laumont and colleagues demonstrated that in murine cancer cell lines and in human primary tumors, 90% of the identified tumor-associated antigens had originated from non-canonical regions (55, 63). In addition, Chen and colleagues recently demonstrated that 240 non-canonical peptides derived from upstream open reading frames located in the 5’UTR and long non-coding RNAs of extragenic DNA were presented by the HLA of human tumor cell lines (62). This new knowledge would not have been generated if the tumor associated antigens were identified by standard exome-based approaches. A recent work by Cleyle and colleagues demonstrated the presence of MAPs originating from non-coding regions in MSS and MSI CRCs (85). However, it remains largely unknown whether tumor specific antigens loaded on the MHC class I can trigger an immune response (85). Current human models do not allow to determine whether MAPs can be edited by the immune system of the host. To bridge this gap, we studied the contribution of tumor associated antigens originating from non-canonical genome in a MMRd murine cell line and its isogenic MMRp counterpart. We performed high depth WGS of *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ isogenic CT26 cells and observed an increase of the mutational burden associated with mismatch repair inactivation across all the genome and particularly in the extragenic and intronic portions. Then, we injected both isogenic cell lines into immunocompromised and immunocompetent mice and used WGS to establish the mutational burden and to identify the genome areas poorly represented after *in vivo* growth which we considered as evidence that immune editing had occurred. Interestingly, we found a significant reduction of alterations in the coding, 5’UTR and 3’UTR regions in *Mlh1*$^{-/-}$ tumors grown in immunocompetent mice. Next, to identify peptides loaded on the MHC class I complex we built an immune-peptidomic pipeline combining RNA sequencing and mass spectrometry technology. Since the identification of amino acid sequences bound to the MHC class I complex requires a list of candidate peptides to be matched with, we assembled two specific RNA databases with all the peptide sequences potentially generated by the transcripts of MMRp and MMRd CT26 cell lines. We specifically selected peptides retained in tumors grown in immunocompromised mice and at the same time lost in immunocompetent mice. To selectively identify peptides originating as a result of a defective MMR system, *Mlh1*$^{+/+}$ sequences were removed from the *Mlh1*$^{-/-}$ database. This approach led the identification of hundreds of MAPs in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ CT26 cells. Finally, to characterize the mutational status and the areas of the genome from which MAPs originated, the sequences obtained from the immune-peptidomic pipeline were combined with the WGS data. Our results show that in both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells most of the MAPs edited in immunocompetent mice originated from non-coding DNA portions in accordance with previous studies (55, 86).

Furthermore, the non-mutated MAPs targeted by the immune system in *Mlh1*$^{-/-}$ predominantly originated from non-coding regions whereas mutant MAPs derived primarily from the UTR and coding regions. To define the relative contribution in terms of immunogenicity between non-mutated MAPs and mMAPS, we first calculated their representativeness by the number of supported RNA sequences; then, we calculated the fold change between the number of MAPs lost in tumor cells after *in vivo* growth and those previously present at the day of injection. Interestingly, we observed that the mMAPs were immune edited more than the non-mutated sequences.

A limitation of the present study is that a definitive conclusion cannot be drawn on the efficacy of single peptides in prompting an immune response. Future studies are needed to functionally validate the identified peptides either by vaccination strategy or *in vitro* immune activation assays. In conclusion, we provide functional evidence that non-coding DNA sequences, which represent 98% of the genome, can contribute to the immunogenic features of MMRd tumors. Additionally, our findings support the relevance of a thorough characterization of tumor samples at the genomic levels including the often overlooked “dark” portion of the genome.

An elevated mutational burden is currently considered a promising independent prognostic biomarker for MMRd cancer (87, 88). In addition, many CRC patients display a low TMB, and are not considered good candidates for ICB therapies. However, most TMB analyses are performed by WES or by custom panels that include a limited number of genes or a portion of them. Accordingly, in most of the studies, the extragenic areas of the genome, that are the vast majority of the entire DNA sequence, are not included in the TMB evaluation. This is noteworthy considering recent findings from Frigola and colleagues that demonstrate how generation of mutations occurs at lower levels in coding than in the non-coding regions (89). Notably, they showed that mismatches in exonic DNA are repaired by MMR more efficiently than in their intronic counterparts. Therefore, non-coding regions could accumulate more alterations during tumor evolution as a result of distinct DNA repair efficiency. These findings lead us to speculate that: i) the evaluation of the extragenic part of the genome could improve the definition of the tumor mutational landscape; ii) in MMRd tumors the contribution of extragenic alterations to generating an immune response could be more impactful than the intragenic part, considering the diverse level of fidelity between intra- and extragenic DNA repair pathways.

Conclusions and future perspectives

Our results reveal the importance of evaluating the diversity of neoepitope repertoire in MMRd tumors to better understand the mechanisms behind the immunogenic properties of these tumor types. We showed that alterations in MMR affect both the quantity of mutations and how these are distributed in the entire genome.

Our study also highlights the role of non-canonical MAPs in triggering an immune response in MMRd mouse models. We point out that 5'UTR and 3'UTR regions are a source of mutated peptides that can be loaded on the MHC class I complex. Furthermore, we found that these candidate mMAPs are lost after growth of MMRd CRC tumors in immunocompetent animals whereas they are preserved in immunocompromised mice. These results suggest that non-canonical MAPs are targeted by the immune system of the host contributing to the immune editing process that controls MMRd tumor growth.

We provide a proof-of-concept that in MMRd tumors non-canonical translational events across the entire genome, i.e., translation of non-coding and out-of-frame coding regions, can effectively contribute to the immunogenic properties of these tumor types. Moreover, the action of the immune system against non-coding region derived neoantigens could be more relevant than that against peptides generated from the intragenic part, considering the diverse level of fidelity between intra- and extragenic DNA repair.

Finally, our results pinpoint the contribution of non-canonical neoantigens in the positive outcome of MMR deficient CRC tumors and provide the rationale for exploring the immunogenic contribution of non-coding genome also in the majority of MMR proficient CRC patients that are immune refractory and not eligible for immune-based therapies.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

2. National Cancer Institute Surveillance, Epidemiology, and End Results Program. Cancer stat facts: colorectal cancer. Accessed September, 2022. https://seer.cancer.gov/statfacts/html/colorect.html.

3. Biller LH, Schrag D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. JAMA. 2021;325(7):669-85.

4. Germano G, Amirouchene-Angelozzi N, Rospo G, Bardelli A. The Clinical Impact of the Genomic Landscape of Mismatch Repair-Deficient Cancers. Cancer Discov. 2018;8(12):1518-28.

5. De Palma FDE, D'Argenio V, Pol J, Kroemer G, Maiuri MC, Salvatore F. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. Cancers (Basel). 2019;11(7).

6. Rospo G, Lorenzato A, Amirouchene-Angelozzi N, Magrì A, Cancelliere C, Corti G, et al. Evolving neoantigen profiles in colorectal cancers with DNA repair defects. Genome Med. 2019;11(1):42.

7. Koopman M, Kortman GA, Mekenkamp L, Ligtenberg MJ, Hoogerbrugge N, Antonini NF, et al. Deficient mismatch repair system in patients with sporadic advanced colorectal cancer. Br J Cancer. 2009;100(2):266-73.

8. Germano G, Lamba S, Rospo G, Barault L, Magrì A, Maione F, et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. Nature. 2017;552:116.

9. De Smedt L, Lemahieu J, Palmans S, Govaere O, Tousseyn T, Van Cutsem E, et al. Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. Br J Cancer. 2015;113(3):500-9.

10. Tougeron D, Sueur B, Zaanan A, de la Fouchardiére C, Sefrioui D, Lecomte T, et al. Prognosis and chemosensitivity of deficient MMR phenotype in patients with metastatic colorectal cancer: An AGEO retrospective multicenter study. Int J Cancer. 2020;147(1):285-96.

11. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372(26):2509-20.

12. Overman MJ, Lonardi S, Wong KYM, Lenz HJ, Gelsomino F, Aglietta M, et al. Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer. J Clin Oncol. 2018;36(8):773-9.

13. Amodio V, Mauri G, Reilly NM, Sartore-Bianchi A, Siena S, Bardelli A, et al. Mechanisms of Immune Escape and Resistance to Checkpoint Inhibitor Therapies in Mismatch Repair Deficient Metastatic Colorectal Cancers. Cancers (Basel). 2021;13(11).

14. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol. 2006;7(5):335-46.

15. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010;138(6):2073-87.e3.

16. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.

17. Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell. 2017;168(4):644-56.

18. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101.

19. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52-60.

20. Chung J, Maruvka YE, Sudhaman S, Kelly J, Haradhvala NJ, Bianchi V, et al. DNA Polymerase and Mismatch Repair Exert Distinct Microsatellite Instability Signatures in Normal and Malignant Human Cells. Cancer Discov. 2021;11(5):1176-91.

21. Haradhvala NJ, Kim J, Maruvka YE, Polak P, Rosebrock D, Livitz D, et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat Commun. 2018;9(1):1746.

22. Nakayama M. Antigen Presentation by MHC-Dressed Cells. Front Immunol. 2014;5:672.

23. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. 2017;18(8):1009-21.

24. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017;357(6349):409-13.

25. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348(6230):69-74.

26. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Rep. 2016;15(4):857-65.

27. Gubin MM, Schreiber RD. CANCER. The odds of immunotherapy success. Science. 2015;350(6257):158-9.

28. Maby P, Tougeron D, Hamieh M, Mlecnik B, Kora H, Bindea G, et al. Correlation between Density of CD8+ T-cell Infiltrate in Microsatellite Unstable Colorectal Cancers and Frameshift Mutations: A Rationale for Personalized Immunotherapy. Cancer Res. 2015;75(17):3446-55.

29. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. Cancer Discov. 2015;5(1):43-51.

30. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science. 2006;313(5795):1960-4.

31. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. J Pathol. 2014;232(2):199-209.

32. Argilés G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(10):1291-305.

33. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet. 2018;391(10135):2128-39.

34. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature. 2014;515(7528):577-81.

35. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015;350(6257):207-11.

36. Picard E, Verschoor CP, Ma GW, Pawelec G. Relationships Between Immune Landscapes, Genetic Subtypes and Responses to Immunotherapy in Colorectal Cancer. Front Immunol. 2020;11:369.

37. Weiner LM, Murray JC, Shuptrine CW. Antibody-based immunotherapy of cancer. Cell. 2012;148(6):1081-4.

38. Yarchoan M, Hopkins A, Jaffee EM. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N Engl J Med. 2017;377(25):2500-1.

39. Zhu M, Jin Z, Hubbard JM. Management of Non-Colorectal Digestive Cancers with Microsatellite Instability. Cancers (Basel). 2021;13(4).

40. Chalabi M, Fanchi LF, Dijkstra KK, Van den Berg JG, Aalbers AG, Sikorska K, et al. Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. Nat Med. 2020;26(4):566-76.

41. Cercek A, Lumish M, Sinopoli J, Weiss J, Shia J, Lamendola-Essel M, et al. PD-1 Blockade in Mismatch Repair-Deficient, Locally Advanced Rectal Cancer. N Engl J Med. 2022;386(25):2363-76.

42. Bjerregaard AM, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunol Immunother. 2017;66(9):1123-30.

43. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016;8(1):11.

44. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32(4):511-7.

45. @font-face ReaV, {font-family:&amp, amp, quot, Math&amp C, amp, et al. A computational tool for designing personalized cancer vaccines. bioRxiv; 2018.

46. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. 2019;11(1):56.

47. De Mattos-Arruda L, Vazquez M, Finotello F, Lepore R, Porta E, Hundal J, et al. Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. Ann Oncol. 2020;31(8):978-90.

48. Bauer DC, Zadoorian A, Wilson LOW, Thorne NP, Alliance MGH. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Brief Bioinform. 2018;19(2):179-87.

49. Boegel S, Castle JC, Kodysh J, O'Donnell T, Rubinsteyn A. Bioinformatic methods for cancer neoantigen prediction. Prog Mol Biol Transl Sci. 2019;164:25-60.

50. Vitiello A, Zanetti M. Neoantigen prediction and the need for validation. Nat Biotechnol. 2017;35(9):815-7.

51. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547(7662):217-21.

52. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017;547(7662):222-6.

53. McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016;351(6280):1463-9.

54. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. Cell Mol Life Sci. 2018;75(4):607-21.

55. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure JP, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10(470).

56. Ruiz Cuevas MV, Hardy MP, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. 2021;34(10):108815.

57. Zhao Q, Laverdure JP, Lanoix J, Durette C, Côté C, Bonneil É, et al. Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. Cancer Immunol Res. 2020;8(4):544-55.

58. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol. 2022;40(2):209-17.

59. Dong C, Cesarano A, Bombaci G, Reiter JL, Yu CY, Wang Y, et al. Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. Oncogene. 2021;40(42):6130-8.

60. Zhang Q, Wu E, Tang Y, Cai T, Zhang L, Wang J, et al. Deeply Mining a Universe of Peptides Encoded by Long Noncoding RNAs. Mol Cell Proteomics. 2021;20:100109.

61. Goodenough E, Robinson TM, Zook MB, Flanigan KM, Atkins JF, Howard MT, et al. Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR. Proc Natl Acad Sci U S A. 2014;111(15):5670-5.

62. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367(6482):1140-6.

63. Laumont CM, Daouda T, Laverdure JP, Bonneil É, Caron-Lizotte O, Hardy MP, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun. 2016;7:10238.

64. Kracht MJ, van Lummel M, Nikolic T, Joosten AM, Laban S, van der Slik AR, et al. Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes. Nat Med. 2017;23(4):501-7.

65. Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. Nat Rev Immunol. 2016;16(4):207-19.

66. Probst P, Kopp J, Oxenius A, Colombo MP, Ritz D, Fugmann T, et al. Sarcoma Eradication by Doxorubicin and Targeted TNF Relies upon CD8. Cancer Res. 2017;77(13):3644-54.

67. Prensner JR, Enache OM, Luria V, Krug K, Clauser KR, Dempster JM, et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol. 2021;39(6):697-704.

68. Pritchard AL, Hastie ML, Neller M, Gorman JJ, Schmidt CW, Hayward NK. Exploration of peptides bound to MHC class I molecules in melanoma. Pigment Cell Melanoma Res. 2015;28(3):281-94.

69. Kalaora S, Barnea E, Merhavi-Shoham E, Qutob N, Teer JK, Shimony N, et al. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. Oncotarget. 2016;7(5):5110-7.

70. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367-72.

71. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218-23.

72. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8(5):1365-79.

73. Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, et al. The translation of non-canonical open reading frames controls mucosal immunity. Nature. 2018;564(7736):434-8.

74. Corti G, Bartolini A, Crisafulli G, Novara L, Rospo G, Montone M, et al. A Genomic Analysis Workflow for Colorectal Cancer Precision Oncology. Clin Colorectal Cancer. 2019;18(2):91-101.e3.

75. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

76. Institute B. Picard Toolkit. GitHub Repository2019.

77. Germano G, Lu S, Rospo G, Lamba S, Rousseau B, Fanelli S, et al. CD4 T Cell-Dependent Rejection of Beta-2 Microglobulin Null Mismatch Repair-Deficient Tumors. Cancer Discov. 2021;11(7):1844-59.

78. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat Cancer. 2020;1(2):249-63.

79. Matafora V, Corno A, Ciliberto A, Bachi A. Missing Value Monitoring Enhances the Robustness in Proteomics Quantitation. J Proteome Res. 2017;16(4):1719-27.

80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

81. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38(18):e178.

82. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64.

83. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941-D7.

84. Loeb LA. Human cancers express mutator phenotypes: origin, consequences and targeting. Nat Rev Cancer. 2011;11(6):450-7.

85. Cleyle J, Hardy MP, Minati R, Courcelles M, Durette C, Lanoix J, et al. Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability. Mol Cell Proteomics. 2022;21(5):100228.

86. Ehx G, Larouche JD, Durette C, Laverdure JP, Hesnard L, Vincent K, et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity. 2021;54(4):737-52.e10.

87. Marabelle A, Fakih M, Lopez J, Shah M, Shapira-Frommer R, Nakagawa K, et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. 2020;21(10):1353-65.

88. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019;51(2):202-6.

89. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017;49(12):1684-92.

Some of the figures in this thesis were produced using Biorender®.