Identification of Positively and Negatively Selected Driver Gene Mutations Associated With Colorectal Cancer With Microsatellite Instability

Vincent Jonchere,1,2,* Laetitia Marisa,1,2,* Malorie Greene,1,* Alain Virouleau,1,3,4 Olivier Buhard,1 Romane Bertrand,1 Magali Syrcek,1,5 Pascale Cervera,1,5 Anastasia Goloudina,1,6 Erell Guillem,1,7 Florence Coulet,1,7 Samuel Landman,1 Toky Ratovomanana,1 Sylvie Job,2 Mira Ayadi,2 Nabila Elarouci,2 Lucile Armenoult,2 Fatiha Merabtene,1,8 Sylvie Dumont,1,8 Yann Parc,1,9 Jérémie H. Lefèvre,1,9 Thierry André,1,10 Jean-François Fléjou,1,5 Agathe Guilloux,3,4,§ Ada Collura,1,§ Aurélien de Reyniès,2,§ and Alex Duval1,§

1Sorbonne Université, University Pierre and Marie CURIE Paris 06, INSERM, Unité Mixte de Recherche938, Equipe Instabilité des Microsatellites et Cancer, Centre de Recherche Saint Antoine, Paris, France; 2Programme Cartes d’Identité des Tumeurs, Ligue Nationale Contre Le Cancer, Paris, France; 3Laboratoire de Mathématiques et Modélisation d’Évry, University Évry, Évry, France; 4Centre National de la Recherche Scientifique, Université Paris-Saclay, Évry, France; 5Service d’Anatomie et Cytologie Pathologiques, Assistance publique - Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France; 6Inovarion, Collaborative research Department Paris, France; 7Genetics Department, Assistance publique - Hôpitaux de Paris, Pitié Salpêtrière Hôpital, Paris, France; 8University Pierre and Marie CURIE Paris 06, Unité Mixte de Service 30 L’Unité Mixte de service Imagerie Cytométrie, Plateforme d’Histomorphologie, Sorbonne Université Paris, France; 9Service de Chirurgie Générale et Digestive, Assistance publique - Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France; 10Department of Oncology, Assistance publique - Hôpitaux de Paris, Hôpital Saint Antoine, Paris, France

**SUMMARY**

Recent studies have shown that cancers arise as a result of the positive selection of driver somatic events in tumor DNA, with negative selection playing only a minor role, if any. The present work indicates that in microsatellite instability cancer, the high level of genomic instability generates both positively selected somatic mutations that contribute to the tumorigenic process but also recurrent somatic mutational events that are negatively selected due to their deleterious for the tumor cells.

**BACKGROUND & AIMS:** Recent studies have shown that cancers arise as a result of the positive selection of driver somatic events in tumor DNA, with negative selection playing only a minor role, if any. However, these investigations were concerned with alterations at nonrepetitive sequences and did not take into account mutations in repetitive sequences that have very high pathophysiological relevance in the tumors showing microsatellite instability (MSI) resulting from mismatch repair deficiency investigated in the present study.

**METHODS:** We performed whole-exome sequencing of 47 MSI colorectal cancers (CRCs) and confirmed results in an
A cquisition of the multiple hallmarks of cancer mainly is owing to somatic mutations. These hallmarks are a convenient organizing principle to rationalize the growth and complexity of tumors (for review, see Hanahan and Weinberg ). Underlying these mutations is the characteristic of genomic instability. This leads to the generation of mutant genotypes that confer advantages or disadvantages to the cells in which they occur; thus allowing the cells to dominate or to involute within the tumor mass. Data obtained from the analysis of thousands of tumors from different primary sites have shown that unlike species evolution, positive selection outweighed the negative selection of somatic mutational events during tumor progression. Different types of genomic instabilities have been described in human malignancies, including a subset of cancers that is characterized by inactivating alterations of mismatch repair (MMR) genes. These tumors show a distinctive phenotype referred to as microsatellite instability (MSI). MSI affects thousands of microsatellite DNA sequences, although numerous alterations also occur in nonrepetitive DNA sequences during tumor progression. This phenotype was first observed in tumors from individuals with the familial cancer condition known as Lynch syndrome, and later in sporadic colon, gastric, endometrial, and other cancer types. The activating BRAF V600E somatic hotspot mutation, affecting a nonrepetitive coding DNA sequence, plays an important role in the progression of sporadic MSI colorectal cancer (CRC). However, most somatic mutations with a postulated role in MSI tumorigenesis are found in microsatellites contained within coding regions, and to a much lesser extent in microsatellites contained within noncoding gene regions (eg, intronic splicing, or in the 5′ UTR or 3′ UTR). Because microsatellites constitute hot spots for mutations in MSI tumors regardless of their location in genes and the function of these genes, such frequent mutations could be neutral or even detrimental to tumorigenesis. In accordance with this working hypothesis, we previously reported frequent inactivation of the HSP110 oncogenic chaperone in MSI CRC.

Recent advances in high-throughput sequencing have made it possible to identify all genetic changes in human MSI neoplasms. Kim et al reported a global view in 27 colon and 30 endometrial tumors with the MSI phenotype. With regard to the selection of MSI-driven events, these investigators did not take into account the strong influence of the length and nature of DNA repeats on the frequency of their instability, as shown earlier by several groups. Furthermore, nucleotide instability outside of DNA microsatellites was not investigated, even though this is an important part of the landscape of somatic changes in MSI CRC. Other studies have attempted to identify driver genes containing selected mutations, or to use various probabilistic models of unselected mutations in MSI CRC while ignoring negative selection, which is more difficult to establish. A recent study reported that tumors with a mutator phenotype (including MMR-deficient cancers) acquired more positively selected driver mutations than other tumors, but found no evidence of negative selection. The latter study investigated substitutions at nonrepetitive sequences, without taking into account repetitive sequences that have high physiopathologic relevance in these tumors.

In the present study we performed whole-exome sequencing (WES) of 47 MSI CRCs and validated results in an independent series of 53 MSI CRCs from the The Cancer Genome Atlas (TCGA). Overall, our results shed new light on MMR-deficient tumorigenesis and suggest that genomic instability in MSI CRC plays a dual role in achieving tumor cell
transformation. They show hitherto unknown pathophysiological aspects of MSI colon tumors that could lead to novel therapeutic approaches specific for this tumor subtype.

Materials and Methods

**Tumor Cohort of MSI CRC Patients Analyzed by WES**

Forty-seven patients who underwent surgical resection for MSI CRC from the Hôpital Saint Antoine (Paris, France) were selected for this study. Tumor samples and adjacent normal tissue counterparts were collected and stored frozen at -80°C before DNA extraction. DNA was purified using the Qiamp protocol (Qiagen, Courtaboeuf, France) as recommended by the manufacturer. Informed consent was obtained for all patients. Gene expression for 30 samples from this cohort was analyzed previously on the Affymetrix U133 plus 2 chips as described (Thermo Fisher Scientific, Waltham, MA).23

**Tumor Cohort of MSI CRC Patients for Survival Analysis**

A total of 164 MSI CRC samples with available whole-genome amplification (WGA) DNA were analyzed further for association between MSI mutational events and relapse-free survival (RFS).16

**Exome Data Analyses**

**WES.** For the 47 pairs of MSI CRC and paired adjacent normal mucosa, 3 µg of genomic DNA was fragmented by sonication and purified to obtain fragments of 150 to 200 base pairs (bp). The oligonucleotide adapters for sequencing were ligated to DNA fragments and purified. After purification, exonic sequences were captured by hybridizing the sequences to biotinylated exon library baits, which were then captured with streptavidin-coated magnetic beads that complex with biotin (SureSelect Human All Exon Kit v5+UTR, 75 Mb; Agilent, Les Ulis, France). The eluted fraction then was amplified by 4–6 polymerase chain reaction (PCR) cycles and sequenced on an Illumina HiSeq 2000 sequencer as paired-end 75 bp (San Diego, CA). Image analysis and base calling were performed using the Illumina Real-Time Analysis Pipeline version 1.14 with default parameters. Read-sequence Fastq files were generated and quality control was checked following Illumina’s recommendations and FastQC reports.

**Overall mutation (single-nucleotide variant and insertion/deletion) calling.** The exome data analysis was first performed using Illumina CASAVA 1.8.2 software (San Diego, CA), which includes reads mapping and variant calling. Exome sequencing data have been deposited in the European genome-phenome archive (accession: EGAS00001002477). Reads were aligned against the hg19 genome build (GRCh37 - Human genome assembly 19) with ELANDv2, a gapped and multisited aligner that reduces artifactual mismatches and allows the identification of small insertions/deletions (indels) (≤10 nucleotide), which is mandatory for analyzing microsatellite instability. Casava then detects single-nucleotide variants and indel variants independently in the tumor and normal samples. To distinguish somatic from germline variants, the results are combined and a Fisher test for base distribution between normal and tumor DNA is computed. We then applied a previously described method24 to generate a list of somatic variants. Quality control filtering removed variants sequenced in <10 reads, with <3 variant calls or with a Quality-Phred of <20. Variants were considered to be of somatic origin when the frequency of variant reads was ≥10% in the tumor and ≤5% in the normal counterpart, with significant enrichment of variant calls in the tumor as assessed by the Fisher exact test (P < .05). Variants then were annotated with Annovar for gene symbol, gene structure location, and exonic functional impact using the RefGene database (hg19 version). Common polymorphisms with a reported frequency of >1% were removed after comparison with the 1000 Genomes Project database and a proprietary database of exomes from normal tissues. Variants were functionally annotated using the annotations provided by Annovar based on the LJB databases (include SIFT scores, PolyPhen2 HDIV scores, PolyPhen2 HVAR scores, LRT scores, MutationTaster scores, MutationAssessor score, FATHMM scores, GERP++ scores, PhyloP scores, and SiPhy scores). When the annotation was not provided, PolyPhen2, Sift, and Provean software were used. If at least one of the methods defined the mutation as damaging, the mutation was annotated as deleterious. The mutation incidence in each tumor was evaluated by dividing the number of somatic mutations by the number of exonic bases covered by ≥10× in both the tumor and normal samples. Mutations were classified into nonrepetitive (NR) and repetitive (R) sequences using the microsatellite list defined with MSIsensor (see later).

**Mutation calling in microsatellite sequences.** To analyze mutations extensively at microsatellite sequence sites, we used the software MSIsensor25 version 0.2, a program for detecting somatic microsatellite changes. First, the list of microsatellites was generated using the scan command of MSIsensor, which searches for sequences of 1–5 bases repeated at least 5 times in the human reference genome sequence (NCBI build37.1 genome fasta file). Then, using the MSIsensor msi command, the mutation status for each microsatellite site (with ≥20 mapped reads) and each tumor/normal tissue pair was estimated by comparing the read-length distribution between tumor and normal samples using the chi-square test. P values for each microsatellite were extracted from MSIsensor outputs and used to define the microsatellite mutation status (if P < .05, the microsatellite was considered to be mutated). Each microsatellite was annotated for gene symbol and gene region type location (exonic, intronic, 5’ UTR, and 3’ UTR) using Annovar, according to the RefGene database (hg19 version).

**Driver mutation selection in nonrepetitive sequences.** Casava mutation (SNV and indel) calling results in coding regions was defined by Annovar and outside repetitive sequences. We applied MutSigCV (v1.4) and Intogen online software with default parameters. Many genes previously published as potential drivers in MSI CRC were not found with those 2 gold standard methods. To allow
recovering these genes, we implemented a simpler method: intrasample binomial laws are fitted to report the different probabilities of mutation from one sample to another; given a sequence, its probability to be mutated at least once is calculated in each sample according to the corresponding binomial law; these probabilities then are combined across samples using the Fisher combined probability test. The resulting statistics then are compared with an empiric null distribution drawn using intronic and synonymous mutations. **Driver mutation selection in repetitive sequences.** For each repeat $i = 1, \ldots, n$ ($7 \times 10^5 \leq n \leq 3 \times 10^6$), we observed the following: a count $y_i$ with values in $[0, n_i]$, where $n_i$ is the number of observed tumor samples for this repeat, and the repeat length $x_i$ (with values from 5 to 27). We tried logistic models on the data but they miss fitting our data well enough for 2 reasons: the observed mutation proportion for 1 repeat never goes to 1, and we observed overdispersion as compared with the logistic model. We consequently added the parameter $\alpha$ to adjust for the first observation and chose a 2-layer-model to model the overdispersion.\(^{29}\) The model contained 2 layers. In the first layer, let $\pi_i = \frac{\exp(\alpha + \beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$ and $\tilde{\pi}_i \sim \beta(\pi_i, (1 - \pi_i)c)$, where $\beta$ stands for the $\beta$ distribution. The random variable $\pi_i$ has moments equal to $E(\tilde{\pi}_i) = \pi_i; \quad \exp(\alpha + \beta_0 + \beta_1 x_i)$ and $\text{Var}(\tilde{\pi}_i) = \pi_i(1 - \pi_i)$ with $c = (\beta^{-1} - 1)$. In the second layer, the distribution of $y_i$ conditionally to $\tilde{\pi}_i$ is $B(n_i, \tilde{\pi}_i)$, the joint distribution of $(Y_i, \tilde{\pi}_i)$ then is proportional to $\tilde{\pi}_i^{y_i}(1 - \tilde{\pi}_i)^{(n_i - y_i)}$ where $\tilde{\pi}_i \sim B(n_i, \tilde{\pi}_i)$, and $V(\tilde{\pi}_i) = \emptyset \tilde{\pi}_i(1 - \pi_i)$ with $c = (\beta^{-1} - 1)$. In the second layer, the distribution of $y_i$ conditionally to $\tilde{\pi}_i$ is $B(n_i, \tilde{\pi}_i)$, the joint distribution of $(Y_i, \tilde{\pi}_i)$ then is proportional to $\tilde{\pi}_i^{y_i}(1 - \tilde{\pi}_i)^{(n_i - y_i)}\tilde{\pi}_i^{c-1}(1 - \tilde{\pi}_i)^{(1 - \pi_i)c - 1)} \frac{1}{B(\pi; (1 - \pi)c)}$, where $B(u, v)$ is the beta function at points $u$ and $v$. The marginal distribution of $y_i$ then is proportional to $B(y_i + n_i - 1, c) \frac{Y_i^{y_i}(1 - Y_i)^{(n_i - y_i)}\Gamma(\pi c)(\Gamma(1 - \pi)c)}{\Gamma((1 - \pi)c)}\Gamma((n_i + 1)(1 - \pi)c)}$, where $\Gamma$ is the gamma function. The log-likelihood term \(\log f(y_i; c, \alpha, \beta_0, \beta_1)\) for individual $i$ then equals $(\log(\Gamma(y_i + \pi c)) - \log(\Gamma(\pi c)) + \log(\Gamma(n_i - y_i + (1 - \pi)c) - \log(\Gamma((1 - \pi)c)) - \log(\Gamma(n_i + 1)c)) - \log(\Gamma(c))$ up to a constant (depending on $y_i$ and $n_i$).

To take into account the presence of outliers in the data, a parameter $\nu_i$ was added in the probability $\pi_i$, in the same fashion as described by Tibshirani and co-workers,\(^{27}\) which becomes $E(\tilde{\pi}_i) = \frac{\exp(\alpha + \beta_0 + \beta_1 x_i + \nu_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \nu_i)}$. To overcome the high dimensionality of our model, we added a lasso penalty.\(^{27}\) The optimization problem to solve is as follows: $\min_{\alpha, \beta_0, \beta_1, \nu_i} -\sum_{i=1}^{n} \log f(y_i; c, \alpha, \beta_0, \beta_1, \nu_i) + \lambda\|\nu\|_1$. It is performed through a proximal-stochastic gradient descent (see Bertsekas\(^{28}\)). All algorithms were developed in Python 3.5 and are available on request. The regularization parameter $\lambda$ in the lasso usually is chosen through cross-validation. However, in our context, this technique does not seem relevant. Hence, cross-validation techniques split the data in training and testing sets, fitting the model on the training set and measure its performance on the testing set, across a grid of tuning parameters to choose the best. The problem here is that we obviously cannot learn the outlyingness of an observation thanks to other observations. In other words, what we learn in the training parts is not useful for the testing parts. Moreover, the result of the cross-validation would highly depend on how the outliers are distributed in the sets. Therefore, we used a heuristic technique of the L-curve, allowing us to choose a $\lambda$, providing us with a high likelihood together with control of the number of outliers. Once the hyperparameters $\alpha, \beta_0$ and $\beta_1$ were estimated, Pearson residuals\(^{29}\) were calculated for each repeat. We then applied the Benjamini Hochberg Yekutieli multistest procedure to detect significantly large residuals. Repeats with detected residuals were finally assigned to the positively selected, respectively the negatively selected when their associated residuals were negative, respectively positive.

**Functional analysis in CRC cell lines and primary colon tumor samples.** To analyze for the enrichment of genes belonging to specific biological processes (Gene ontology), mutated genes that were positively or negatively selected were analyzed using DAVID\(^{30}\) against the Homo sapiens database ($P < .05$; number of genes, $\geq 5$).

CRC cell lines were purchased from the American Type Culture Collection (Manassas, VA). All cells were cultured in Dulbecco’s modified Eagle medium supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin in a humidified atmosphere at 37°C supplied with 5% CO$_2$. All cell lines were mycoplasma free. Primary tumors and normal colon tissues were obtained from patients with CRC undergoing surgery at the Hospital Saint-Antoine between 2009 and 2014 and after informed patient consent was obtained and approval from the Institutional Review Boards/Ethics Committees of Hospital Saint-Antoine (Paris, France). Patients with CRC (1998–2007) from 6 centers involved in a study of MSI status were described previously.

### Mutation Analysis
Specific primers for exonic coding of DNA repeats of negatively selected gene mutations were designed using AmpliFlix software (V1.7). PCR amplification was performed on tumor DNA amplified by WGA technology using the Illustra GenomiPhi DNA Amplification V2 kit (GE Healthcare, Velizy-Villacoublay, France). Absence of artifactual alteration of microsatellite sequences caused by WGA was validated by comparing Amplified Fragment-Length Polymorphism traces of several long microsatellites before and after WGA on a 3100 GA (Applied Biosystems, Foster City, CA). Fluorescent PCR products were run on an ABI PRISM 3100 Genetic Analyzer with GS400HD ROX size standard and POP-6 polymer (Thermofisher Scientific, Waltham, MA), and Gene mapper software (V4.0, Thermofisher Scientific) was used to analyze negatively selected mutations in exonic microsatellite traces (Applied Biosystems, Foster City, CA). Oligonucleotide sequences are available on request.

### Transient Gene Silencing by Cell Transfection and Treatments
A total of $1.25 \times 10^5$ cells were cultured in a 6-well plate for 24 hours. Cells then were transfected with Silencer Select (Thermofisher Scientific) small interfering RNAs (siRNAs) (2 targets per gene) using Lipofectamine RNAiMax according to the manufacturer’s recommendations.
(Thermofisher Scientific). siRNA inhibition was assessed 48 hours after transfection by real-time quantitative PCR (Thermofisher Scientific). To induce apoptosis, 48 hours after transfection the cells were treated with TRAIL agent (Invitrogen, Carlsbad, CA, États-Unis) for 3 hours at 50 ng/mL (FET) or for 4 hours at 30 ng/mL (HCT116), 100 ng/mL (SW480, RKO, and SW620), and incubated at 37°C, 5% CO2.

Analysis of Cell Apoptosis
Apoptosis was analyzed by flow cytometry using an Annexin V-fluorescein isothiocyanate and 7-aminoactinomycin D staining kit (Beckman Coulter, Inc, Brea, CA) 48 hours after transfection. Cells were detached using StemPro Accutase cell dissociation reagent at room temperature for 10 minutes and stained with reagents according to the manufacturer’s recommendations. Each sample was evaluated by flow cytometry (Gallios, Beckman Coulter, Inc). Data were analyzed using Kaluza Flow Analysis Software (Beckman Coulter, Inc).

Real-Time Cell Proliferation, Migration Monitoring, and Data Analysis
HCT116 cells were seeded at a density of $3 \times 10^4$ cells/well into E-plate 16 (ACEA Biosciences, Inc, San Diego, CA) and monitored on the xCELLigence Real-Time Cell Analyzer (RTCA) Dual Plate instrument (ACEA Biosciences, Inc) according to the manufacturer’s instructions. Cell proliferation was assessed by electrodes in chambers and impedance differences within an electrical circuit were monitored by the RTCA system every 15 minutes for up to 50 hours. Cell migration was assessed using a CIM plate device of the xCELLigence system. The CIM plate consists of 2 chambers separated by a microporous membrane (pore size, 8 μm) attached to microelectrodes. In this case, the cell index calculated on the basis of impedance measurements reflects the number of cells that migrate through micro pores monitored by the RTCA system every 15 minutes for up to 50 hours. These differences are converted into a cell index. The baseline cell index is determined by subtracting the cell index for a cell-containing well from the cell index of a well with only culture media. The experiments were supervised by an author (A.C.) and all efforts were made to minimize suffering.

Immunohistochemistry
Briefly, 4-μm-thick sections of paraffin-embedded tissue samples were cut onto silane-treated Super Frost slides (CML, Nemours, France) and left to dry at 37°C overnight. Tumor sections were deparaffinized in xylene and rehydrated in pure ethanol. Before immunostaining, antigen retrieval was performed by immersing sections in citrate buffer, pH 6.0 (WNK1) (15 min at 95°C), washed twice in phosphate-buffered saline (PBS) for 3 minutes, and treated with 3% H2O2-PBS for 15 minutes to inhibit endogenous peroxidases. After washing in PBS, slides were saturated for 25 minutes in 3% bovine serum albumin PBS. Sections then were incubated for 1 hour at room temperature with antibody to WNK1 (dilution 1/100; clone ab128858; Abcam, Cambridge, United Kingdom). After washing in PBS, secondary antibody (8114P; Cell Signaling, Danvers, MA) was added for 30 minutes at room temperature. Slides were washed twice for 5 minutes in PBS and shown using the Novared kit (Vector, Burlingame, CA). Slides were washed twice in water for 5 minutes and counterstained with 10% Meyer’s hematoxylin. After 1 wash in water, slides were dehydrated in 100% ethanol and in xylene for 30 seconds each. Apoptosis was quantified by counting the number of labeled cells with anti-caspase 3 antibody per 100 tumor cells in the most affected areas.

2018 The Dual Role of MSI in Colorectal Cancer 281
A. Repartition of R & NR sequences
- NR: 97.3% (7.3 Mb)
- R: 2.7% (2.7 Mb)

B. Mutation effect on exonic sequence
- Deletion: 79.2% (14641/18650)
- Insertion: 12.8% (2255/18650)
- Nonsynonymous: 61.8% (6449/1012034)
- Other: 9.5% (1134/12034)

C. Repartition of R & NR mutations
- NR: 40.2%
- R: 59.8%

D. Stade TNM

E. MSS TCGA (n=293) vs MSI (n=47)

F. Gene mutation frequency
- Adenosine
- Thymine
- Guanine
- Cytosine

Repeat length
Survival Analysis

In the cohort of 164 MSI CRC patients, the association between mutations and survival was assessed by multivariate Cox proportional-hazards regression analyses and adjusted by TNM stage. This was performed for 5 negatively selected target MSI mutations. This also was performed for a Boolean mutational index that was calculated from the mutational status of the 5 target genes in each tumor sample. The proportional-hazards assumption was tested using the cox.zph function. RFS was used and defined as the time from diagnosis to first relapse time or death from a cancer cause only. The cut-off point for statistical significance was .05.

Results

Exome-Wide Analysis of MMR-Deficient CRC: Genomic Instability at Nonrepetitive and Repetitive DNA Sequences

We examined WES data from 47 primary MMR-deficient CRCs defined as having MSI according to international criteria. The genome fraction covered by WES was 75 MB, including UTR (37%), coding exonic (56%), and intronic (7%) regions. Repetitive DNA sequences represent less than 3% of the genome fraction covered by WES (roughly 2 Mb of 75 Mb), with 56% in intronic, 19% in coding exonic, and 25% in UTR regions (Figure 1A). Computational methods were used to

Figure 2. NR mutation types. (A) Distribution of mutation substitution types within the sample. (B) Average proportion across the sample of each transition type. (C) Frequency of C to T mutations according to the flanking base. (D) Distribution of the number of mutations according to the flanking bases for each type of base substitutions.

Survival Analysis

In the cohort of 164 MSI CRC patients, the association between mutations and survival was assessed by multivariate Cox proportional-hazards regression analyses and adjusted by TNM stage. This was performed for 5 negatively selected target MSI mutations. This also was performed for a Boolean mutational index that was calculated from the mutational status of the 5 target genes in each tumor sample. The proportional-hazards assumption was tested using the cox.zph function. RFS was used and defined as the time from diagnosis to first relapse time or death from a cancer cause only. The cut-off point for statistical significance was .05.

Results

Exome-Wide Analysis of MMR-Deficient CRC: Genomic Instability at Nonrepetitive and Repetitive DNA Sequences

We examined WES data from 47 primary MMR-deficient CRCs defined as having MSI according to international criteria. The genome fraction covered by WES was 75 MB, including UTR (37%), coding exonic (56%), and intronic (7%) regions. Repetitive DNA sequences represent less than 3% of the genome fraction covered by WES (roughly 2 Mb of 75 Mb), with 56% in intronic, 19% in coding exonic, and 25% in UTR regions (Figure 1A). Computational methods were used to
identify somatic mutation events at both NR and R DNA sequences (see the Methods section for further details). Investigations were restricted to mononucleotide R sequences because these are the most frequently affected by somatic mutations in MMR-deficient cells and often are located in coding regions or in noncoding UTR sequences endowed with putative functional activity. Repeats of at least 5 nucleotides in length were considered, in accordance with the definition of DNA microsatellite sequences. Mutations in R and NR sequences occurred in similar proportions, representing on average 60% and 40% of all somatic events, respectively (Figure 1B). These mutations accumulated in parallel in MSI tumor samples (Figure 1B) \( P = 2.47 \times 10^{-14} \), \( R = 0.85 \). Relative to the fraction of covered genome, the mutation rate observed in R sequences was approximately 24-fold higher than in NR sequences, expectedly. No significant differences were observed between MSI CRC from Lynch syndrome and sporadic cases, or between tumors with different TNM stages (Figure 1B). A much higher number of mutations was observed in this MSI colon tumor cohort compared with the overall incidence of mutations reported for all CRCs (Figure 1C). This high rate of mutation resulted in a much higher proportion of genes with mutations in coding regions (6% vs 1%) (Figure 1D).

Mutation frequencies also were evaluated in coding, UTR, or intronic regions (Figure 1E and F). For both NR and R sequences, a significantly higher mutation frequency was observed in intronic compared with coding exonic and UTR regions of the tumor genome (Figure 1E), expectedly. We confirmed at the exome scale that both the length and composition amino acid constitution of these DNA repeats (A/T vs C/G) determine their mutational frequency (Figure 1F). There was almost 100% probability of mutation if the microsatellite repeat length was longer than 14 bp, consistent with previous results. Mutation events also were more frequent in G/C nucleotide repeats compared with A/T repeats. These observations indicated that distinct models are needed to analyze the occurrence of mutations in R and NR sequences in MMR-deficient tumors.

Figure 3. Identification of candidate driver genes with mutations in NR sequences. (A) Distribution of mutation types in coding NR regions according to their functional impact (annotation tool Annovar). The functional impact was based on several methods, with a mutation considered as deleterious if at least 1 of those methods estimated it to be deleterious. The average percentage of each mutation type per sample is shown in the inset. (B) Schematic representation of the 3 methods (MutSigCV, Intogen, combined binomial) used in this study to identify driver mutations (see the Methods section for details). (C) Heatmap of the significance (q values) of mutated genes commonly described for their functional impact in MSI CRC using Intogen, combined binomial, and MutSigCV analyses. (D) Oncoprint representation of mutations in coding NR sequences within each sample for the 25 top significantly mutated genes (indicated in darker grey for each approach on the side annotation heatmap) and for the 5 significant genes considered as driver genes by Intogen (indicated in black on the side annotation heatmap) across samples. Top and side bar plots indicate the percentage of each type of mutation within the sample and within the gene, respectively.
Modeling the Occurrence of Mutations in Coding, Nonrepetitive DNA Sequences Identifies Known and New Actors in MSI Colorectal Tumorigenesis

As shown in Figure 1, MSI colon tumors accumulate somatic mutations in R and NR DNA sequences at similar proportions. For NR sequences, the analysis of nucleotide substitutions in MSI tumors is shown in Figure 2. In coding sequences, they mostly consisted of nonsynonymous nucleotide substitutions that probably were deleterious in the majority of cases (>50%) (Figure 3A). Only a small number of events in NR

Figure 4. Positive and negative selection in repetitive sequences in MSI CRC. (A) Distribution model of mutation frequencies across samples in microsatellites in UTRs or coding exonic regions according to repeat length for A/T nucleotides. The color gradient indicates the density of the β-binomial logistic regression model. The blue curve represents the median of observed mutation frequencies, and the red curve represents the median obtained from the model. This figure shows the statistical model and outliers for adenosine/thymine. (B) Box plot representation of mutation frequency variations according to the nucleotide composition of the microsatellite and to the repeat length. The significant independence of the chi-squared distribution is annotated by asterisks as follows: *P < .05, and **P < .01. (C) Distribution of microsatellite mutations (log10 scale) in the 3 gene regions (UTRs, coding exonic, and intronic) according to repeat length. (D) Distribution of outlier mutation in microsatellites contained in UTRs and in coding exons. Positively and negatively selected microsatellite mutations are represented above and below the dotted line, respectively. (E) Distribution of the percentage of outlier mutations (log10 scale) according to repeat length. The significant independence of the chi-squared distribution is annotated by asterisks, as follows: *P < .05, **P < .01, and ***P < .001.
sequences were indels (Figure 3A), in line with a previous report.35

We next aimed to identify mutational events in NR sequences that showed an abnormally high frequency in tumor DNA (ie, positively selected mutational events) (Figure 3B and see the Methods section for details). Overall, we identified the 141 most consensual driver genes of colon tumorigenesis (Figure 3C and Supplementary Table 1). The top 30 driver mutated gene list is shown in Figure 3D and Supplementary Table 2, and includes recognized master genes in colorectal oncogenesis such as \textit{BRAF}, \textit{APC}, \textit{KRAS}, and \textit{PIK3CA} (Figure 3C).

\textbf{Modeling the Occurrence of Mutations in Repetitive Sequences Shows Positively and Negatively Selected Events in the MSI Tumor Genome}

In cancer cells, somatic mutations occur randomly (mutational background) across the genome. In MSI tumor cells, DNA microsatellites constitute natural hot spots for these somatic events and the MSI tumor type is characterized by a high background of instability in repeat sequences. The mutability of DNA repeats within MSI tumors depends on functional and structural factors, as well perhaps on other as yet unidentified factors. Taking into account these previously described structural criteria,13 that is, repeat length and nucleotide composition (adenosine/thymine vs guanine/cytosine) (Figure 1), we developed a statistical model (see Methods for further details) that discriminated 3 functional categories of MSI-linked somatic mutations occurring at DNA repeats: the first category is positively selected events that confer benefits to the MSI tumor cells in which they occur because they have an oncogenic impact. These are believed to be positive drivers of the MSI-driven tumorigenic process and their mutation frequencies are higher than expected by chance in the model. The second category is negatively selected events that are deleterious for the tumor cells in which they occur because they have an anticancer impact. These are believed to be negative drivers of the MSI-driven tumorigenic process and their mutation frequencies are lower than expected by chance in the model. The third category is MSI-linked mutational events owing to background that do not confer benefits or have any oncogenic impact. Their mutation frequencies are found within the background level for MSI by the model. Although such neutral events are not thought to play a role during tumor progression, some gene alterations could have functional significance when they occur together.

Most allelic shifts were deletions and/or insertions of 1 bp, or more rarely 2 bp (data not shown). These were considered equally as mutant alleles in the genomic analysis of instability at mononucleotide repeats. A similar pattern was observed for dinucleotide repeats (data not shown). To build the model, only MSI-related events that occurred in mononucleotide R sequences were considered because these largely predominate over others such as in dinucleotide repeats8,12 (see the Methods section for further details). Repeat length was used as an input parameter for the model and 2 models were fitted: one for A/T composition and the other for G/C composition (Figures 4B and 5). The density of the model for A/T
composition of repeats is shown in Figure 4A. Microsatellites that were shown within our model, abnormally high or low mutation frequency within UTRs, or coding exonic regions are indicated (Figure 4A and Supplementary Table 3). Overall, we identified 1050 and 561 outlier events showing aberrant positive and negative selection in MSI CRC, respectively. These included 1376 mutations in UTR sequences (828 and 548 showing positive or negative selection, respectively). With the exception of these 13 frameshift mutations that affected coding DNA sequences (see later), negatively selected events were observed almost exclusively in noncoding microsatellites (UTRs). In contrast, positively selected mutations were observed in both coding and noncoding DNA repeats (Figure 4C–E).

According to our model, we could only identify negative selection at long DNA repeats, that is, those at least 9 bp in length (Figures 4A and 5). Because these long DNA
### A

#### Negative Selection

| p-value | UTRs, Coding | Coding |
|---------|--------------|--------|
| > 0.05  | R            | R      |
| 0.01    | R            | R      |
| < 0.01  | R            | R      |

### Positive Selection

| UTRs, Coding | Coding |
|--------------|--------|
| R            | R      |

#### Gene Sets

- **Cell communication**
  - TGF-β receptor signaling pathway
  - Wnt receptor signaling pathway

- **Cell differentiation**
  - Epithelial cell differentiation

- **Cell growth and death**
  - Apoptosis
  - Cell proliferation
  - Mitotic cell cycle

- **Cell mobility**
  - Cell adhesion
  - Cell migration
  - Cytoskeleton organization

- **Development/Morphogenesis**
  - Blood vessel development
  - Cell morphogenesis
  - Epithelial tube morphogenesis

- **Immunity**
  - Antigen signaling
  - Lymphocyte activation
  - Lymphocyte differentiation

- **Regulation of transcription**
  - Chromatin organization

- **Metabolism process**
  - Carbohydrate metabolism
  - Protein metabolism
  - RNA metabolism

- **Response to stress**
  - Response to hypoxia

- **Homeostasis process**

### B

#### Ag presentation

Cytokine production

- CD4+ T cell
- CD8+ T cell

- Proteosome

- HLA-DR

- MHC class I

- MHC class II

#### p53/apoptosis

- p53
- Apoptosis

#### MAPK

- ERK
- JNK
- p38

#### Wnt/Wingless

- Wnt
- Wnt3a

- Wnt1

- APC
- β-catenin
- TCF-4

- Frizzled

- Axin

- Notch

- Catenin

- β-catenin binding

- Cell cycle arrest

### C

- **Up-regulation**: Green
- **Down-regulation**: Red

### D

#### Selection

- **Positive**
  - TGF signaling
  - DNA mismatch repair
  - Tumor suppressor
  - Induction apoptosis
  - TGF signaling

- **Negative**
  - Inhibition proliferation
  - Cell proliferation
  - Cell migration

#### Functional annotation

- **Activation of tumor suppressor function**
  - CUL5
  - UTRs
  - UTRs

- **Activation of oncogene function**
  - IL1RN
  - UTRs
  - UTRs

- **Loss of oncogene function**
  - AR
  - UTRs/UTRs
  - UTRs

- **Loss of tumor suppressor function**
  - SMAD4
  - UTRs/UTRs
  - UTRs
repeats are mainly noncoding and located in UTR parts of human genes (5404 UTR candidates vs 248 in the coding DNA), the great majority of negatively selected mutations consequently were identified within UTR DNA repeats. However, when the number of negatively selected events was normalized by taking into account the overall number of microsatellites in coding and UTR regions (Figure 4C–E), no significant enrichment for negative selection in UTR vs coding repeats was observed in MSI CRC (13 of 267 [4.9%] vs 548 of 7030 [7.9%], respectively). Overall, our analysis of MSI through exome sequencing led us to identify 563 mutations that were negatively selected in MSI CRC, representing <10% of the candidate coding and UTR DNA repeats with a size ≥9 bp as described earlier.

Validation of Exome-Wide Analysis of MSI and Refining the List of MSI Target Genes in MSI CRC

We next compared our results with those of the TCGA consortium, which used MuTect2 caller in 53 MSI CRC. The mutation frequencies observed at microsatellite loci were highly similar in both cohorts (R = 0.86; P < 2.10⁻¹⁶) (Figure 6A). Instability at 9 microsatellite loci also was investigated using PCR and Restriction Fragment Length Polymorphism in an independent cohort of 180 MSI CRCs. By using this manual gold standard method, very similar mutation frequencies were observed for these 9 coding repeat sequences in the 2 cohorts with the 2 methods, including 8 in which we validated the low mutation frequency (Figure 6B).
Figure 9. Exonic outlier mutation genes: mutation positions and negative selection validation. (A) Negatively selected coding exonic outliers identified in this study (repeat length: 10). (B) Schematic structure of wild-type (in grey) or mutant (in black) proteins of the 9 outlier mutation genes in which mutations were negatively selected in MSI tumors. *Genes that are not investigated in the present study because mutations do not match with putative loss of function (RXFP1, SYCP1, RNASEH2B) or because gene silencing was not successful (CHD2). (C) Table of the 5 outlier mutations negatively selected in MSI tumors identified from exome sequencing data and investigated here by functional analysis. Four of the 5 genes have a literature-documented role that supports negative selection of their mutations because of the tumorigenic implication.
A. % Apoptotic cells (AnnexinV)

B. Cell index (AU) vs. Hours

C. Cell index (AU) vs. Migration

D. Tumor growth (mm³) vs. Days

The Dual Role of MSI in Colorectal Cancer
According to the published literature, the majority of known and extensively analyzed target gene mutations in MSI CRCs were found here in MMR-deficient CRCs. These included AXIN2, CDX2, BCL10, APAFI, CHK1, PLH3, BLM, RAD50, WIP3, MBD4, CASP5, and AIM2 (Figure 6C). However, TGFBR2, ACVR2A, BAX, MSH3, MSH6, IGF2R, and several others remained in the group of genes with positively selected mutations in MSI tumors. Interestingly, this group mostly contained a small coding repeat (5–7 bp in length) whose mutation frequency was not high but nevertheless was subjected to strong positive selection pressures according to our model (eg, UNC5B, PTEN, and APC). Finally, our signature also contained a small number of target genes with a long coding repeat (9 or 10 bp in length) whose mutations were negatively selected in MSI tumors (Figure 6C) and in which we further assessed the functional impact (see later).

Investigating the Interplay Between MSI, Changes in Gene Expression Level, and Cancer-Related Pathways

We next tested the hypothesis that both positively and negatively selected outlier mutations in R sequences constitute major events in MSI tumorigenesis that result in pro-oncogenic or anti-oncogenic impacts, respectively. To do this we assessed gene ontology terms associated with these mutations and found several to be enriched significantly in such events (Figure 7A and Supplementary Table 4). These outlier mutations were observed in cancer-related pathways known to play an important role in tumor development (eg, Wnt/Wingless and RAF/RAS/MAPK signaling), or with antitumor immunity. Their positive or negative selection in MSI tumors were likely to accord with their expected positive or negative impact, respectively, on the activity of these pathways in CRC (Figures 7B and 8).

We then assessed whether these outlier mutations influenced the expression level of the corresponding target gene in MSI tumors. Several mutations in coding regions and in UTRs were associated with significantly altered gene expression when assessed at the messenger RNA (mRNA) level using transcriptome data from 30 MSI CRC samples (Figure 7C). Because of nonsense-mediated mRNA decay, we mostly observed down-regulation of mutated transcripts from coding regions, as expected. The overall impact of outlier events in UTR tracts was mixed, with down-regulation or up-regulation of a few target genes in MSI CRCs. Based on these results, a list of outlier mutations expected to play an important role in MSI tumor development was proposed (Figure 7D). In line with a protumorigenic effect, positively selected outlier events may inactivate tumor-suppressor functions by down-regulating mRNA expression or activate oncogene functions by up-regulating mRNA expression. Acting in opposition, the negatively selected outlier events could activate tumor-suppressor functions by up-regulating mRNA expression or inactivate oncogene functions by down-regulating mRNA expression, thereby slowing down MSI tumorigenesis.

Functional Validation of the Deleterious Impact of Negatively Selected Coding Mutations on CRC Cells

In MSI tumors, mutations observed in repetitive coding sequences are frameshifts (indels) and generally lead to truncation of the corresponding aberrant protein. Although these events are usually loss-of-function mutations, nonsense-mediated mRNA decay acts to degrade mutant mRNAs that may encode proteins with residual functional activity because these transcripts contain a premature termination codon. We therefore hypothesized that negatively selected mutational events identified in the genomic screen shown in Figure 6 could be deleterious for MSI tumor cells. As stated earlier, only a few of these events were in coding regions and led to truncation of the respective proteins (eg, WNK1, PRRC2C, CHD2, SYCP1, GART, RXXFP2, RFC3, and HMGXB4) (Figure 9). Five of these target genes (WNK1, HMGXB4, PRRC2C, RFC3, and GART) were selected
Figure 2

A) 24H

HCT116

Relative mRNA level (% of expression)

B) 48H

HCT116

Relative mRNA level (% of expression)

SW480

Relative mRNA level (% of expression)

SW620

Relative mRNA level (% of expression)

C) WNK1

Relative mRNA level (% of expression)

GART

Relative mRNA level (% of expression)

HMGXB4

Relative mRNA level (% of expression)

RKO

Relative mRNA level (% of expression)

KM12

Relative mRNA level (% of expression)

SW620

Relative mRNA level (% of expression)

D) RKO

% Apoptotic cells (Annexin V)

km12

% Apoptotic cells (Annexin V)

SW620

% Apoptotic cells (Annexin V)

Untreated

Treated

2018 The Dual Role of MSI in Colorectal Cancer 293
Negatively Selected Events Are Associated With Worse Survival of MSI CRC Patients

We next evaluated whether negatively selected coding sequence mutations that were associated with deleterious effects in CRC cells (eg, microsatellites located in coding regions of *WNK1, HMGXB4, PRRC2C, RFC3*, or *GART*) also may be clinically relevant. An additional cohort of 164 MSI CRC patients originating from 3 clinical centers in France was analyzed by Cox survival models adjusted for TNM stage. In the overall cohort, mutated *WNK1* (hazard ratio [HR], 3.1; 95% CI, 1.2–8; \( P = .02 \)) and *PRRC2C* (HR, 2.9; 95% CI, 1–8.1; \( P = .04 \)) were associated with worse RFS (Figure 14). The *HMGXB4* mutation also showed a trend for association with worse RFS (HR, 2.5; 95% CI, 0.78–7.8; \( P = .12 \)) (Figure 14).

To examine the overall relationship between the 5 negatively selected target gene mutations and patient survival, a mutational index value was computed to summarize this MSI target gene category. Cox modeling based on this representation was associated with significantly worse survival, suggesting an overall negative impact of these mutational events on patient outcome (HR, 3; 95% CI, 1.1–7.9; \( P = .03 \)) (Figure 14).

Discussion

MSI tumors represent a distinctive phenotype characterized by a high background of nucleotidic instability. The present work indicates that, expectedly, in such an MMR-deficient context,\(^{30,13,37}\) genomic instability generates positively selected somatic mutations in both R and NR DNA sequences that are likely to contribute to the tumorigenic process. In addition, it also suggests that MSI tumors must deal with frequent somatic mutational events that are deleterious for the MSI tumor cells and result in a tumor-suppressor effect. Among these mutational events, some should be lethal and therefore not detected in tumor DNA whereas other negatively selected could be deleterious for tumor cells without being lethal, depending on the mutational landscape and other factors. These events represent a weakness of the MSI-driven tumorigenic process. The present results shed new light on MMR-deficient tumorigenesis and suggest that genomic instability in MSI CRC plays a dual role in achieving tumor cell transformation.

Frameshift gene mutations owing to MSI in coding repeats are likely to result in inactivation of the corresponding truncated mutant protein, provided the mutant transcript is not degraded by nonsense-mediated mRNA decay.\(^{10}\) The clearly deleterious consequences of 5 negatively selected coding mutations we report here in both MSI and microsatellite stable tumor cells is of interest. Their frequent somatic inactivation in MSI CRC can impede the progress of cell transformation and lead to the regression of clones in which they occur. A major example of this was *WNK1*, which codes for a positive regulator of canonical Wnt/β-catenin signaling and whose inactivation in different tumor types is deleterious.\(^{38,39}\) Other mutations occurring in *HMGXB4, GART, RFC3, or PRRC2C*, and the silencing of this latter candidate decreased cell proliferation in lung cancer.\(^{44}\) In line with our results, a recent study also found that silencing of some of these targets (*WNK1, RFC3*, and *GART*) was lethal in haploid human tumor cells.\(^{45}\) Although the MMR-deficient tumor cells in which these mutations occurred were eliminated from the bulk of most MSI colon tumors through negative selection, our results also showed that, strikingly, the few tumors in which at least one of these mutations was detected was associated with worse patient prognosis. This suggests the anticancer impact of such mutations should be counterbalanced by other oncogenic processes that remain to be identified and were responsible for the poor prognosis. This clinical observation on patient outcome is intriguing and will require further investigation in larger cohorts. It was difficult to address with the present cohort because of the low

Figure 11. (See previous page). Validation of gene abrogation by the siRNA approach and validation of the deleterious impact of 5 outlier mutations genes in RKO, KM12, FET, and SW620 CRC cell lines. Gene expression (mRNA level) of outlier mutation-related genes after knock-down by single siRNA was assessed at \( (A) \) 24 or \( (B, \text{ left panel}) \) 48 hours after transfection by real-time quantitative PCR. Data represent the means ± SEM of at least 3 independent experiments. \( (B, \text{ right panel}) \) Gene expression (mRNA level) of outlier mutation-related genes after simultaneous down-regulation of 3 genes in HCT116 (upper panel) and SW480 (lower panel) cell lines. \( (C) \) Gene expression (mRNA level) of outlier mutation genes after knock-down by siRNA in 4 cell lines (MSI: RKO, KM12; and microsatellite stable: FET, SW620) was assessed 48 hours after transfection by real-time quantitative PCR. Data represent the means ± SEM of 3 independent experiments. \( (D) \) Flow cytometry analysis of apoptosis (Annexin V) of untreated (triangle) or TRAIL-treated (circle) MSI (RKO and KM12, left panel) and microsatellite stable (FET and SW620, right panel) CRC cell lines transfected either with a single specific siRNA gene (*WNK1, HMGXB4, and/or GART*) or with scrambled siRNA. Data represent the means ± SEM of 3 independent experiments. \( t \) test: \( *P < .05, **P < .01 \) and \( ***P < .001 \) of indicated silencing condition compared with control.
Figure 12. Flow cytometry data. Flow cytometry analysis of early (Annexin V–positive and 7-amino-actinomycin D [7-AAD]–negative cells) and late (Annexin V– and 7-AAD–positive cells) apoptosis of untreated or TRAIL-treated HCT116 (MSI, upper panel) and SW480 (microsatellite stable, lower panel) CRC cell lines transfected either with a single specific siRNA gene (WNK1, HMGXB4, GART, RFC3, and/or PRRC2C) or with scrambled siRNA. One experiment of the 3 performed is shown.
**Figure 13. Experimental data of cell growth and cell migration analysis.** Real-time monitoring of (A) cell growth and (B) cell migration using the xCELLigence system (HCT116 CRC MSI cell line). This system allows estimation of the cell index in real time—the parameter based on impedance measurement and reflecting the number of cells attached to the surface of the experimental chambers. Quadruplicates of 3 independent experiments are shown. (C) H&E stain, and WNK1 marker expression by immunohistochemistry in tumor xenografts at day 30 is shown (see Figure 10D).
frequencies of negatively selected events and the small number of relapses in MSI CRC patients.

Aside from the small number of deleterious mutations found in coding sequences, negatively selected mutational events were found mostly in long noncoding repeats located in the 5′ or 3′ UTR. Although approximately 10% of these somatic mutations were found to alter gene expression at the RNA level, their possible functional impact requires further investigation. We did not perform a functional analysis to show a deleterious impact in MSI CRC cells, as performed for negatively selected events in coding regions. However, these outlier mutations were located in genes with a role in several cancer-related processes, as shown in the pathway enrichment analysis. This indicates their negative selection in MSI CRC cells was not a chance event. Interestingly, some of the negatively selected mutations identified here were found to up-regulate tumor-suppressor functions during MSI tumor development, whereas others were observed to down-regulate oncogene functions. This is in accordance with their paradoxical activation or inactivation, respectively, during the tumorigenic process. Together, these findings highlight that MSI in noncoding UTRs could have an important antitumor impact during MMR-deficient tumor development.

Our analysis of the MSI colon tumor exome confirmed the majority of known target gene mutations for MSI. These and many other mutations in 8- to 10-bp repeats reported previously in the literature are thought to be key events in MSI-driven tumorigenesis (for review see Hamelin et al37). Although these mutations may have functional significance in particular contexts, we showed that the frequency of most of these microsatellite mutations was not different from the background frequency expected for their length, suggesting their overall impact on tumor development may be limited, if any. In contrast, we identified several mutations in smaller coding and noncoding DNA repeats of 5–7 bp in length that showed a high positive selection in MSI CRC. We postulate these new candidate genes for MSI tumor progression that contain relatively short repeats represent important oncogenic driver events in MMR-deficient CRC and may be much more relevant for tumorigenesis than many of the MSI-related mutations reported in the past.

Although almost all sporadic MSI CRC arise because of MLH1 deficiency after epigenetic silencing, Lynch-related MSI CRC is associated with germine mutations in MLH1.

### Table 1. Clinical relevance of MSI-driven coding region mutations in target genes in CRC patients.

| Gene      | n events | p-value | HR (95% CI) | n mut | n wt |
|-----------|----------|---------|-------------|-------|------|
| WNK1      | 22       | 0.024   | 3.1 (1.2 to 8) | 23    | 128  |
| PRRC2C    | 20       | 0.042   | 2.9 (1 to 8.1) | 15    | 120  |
| HMGXB4    | 15       | 0.12    | 2.5 (0.78 to 7.8) | 17    | 86   |
| GART      | 24       | 1       | 1 (0.43 to 2.3) | 43    | 94   |
| RFC3      | 23       | 0.46    | 0.69 (0.25 to 1.8) | 42    | 112  |
| IGF2R     | 11       | 0.3     | 1.9 (0.57 to 6.2) | 35    | 60   |
| TGFBR2    | 12       | 0.76    | 1.3 (0.27 to 6) | 78    | 22   |
| MSH3      | 12       | 0.84    | 0.89 (0.28 to 2.8) | 49    | 53   |
| BAX       | 12       | 0.6     | 0.74 (0.23 to 2.3) | 48    | 52   |
| RAD50     | 12       | 0.27    | 0.51 (0.15 to 1.7) | 50    | 49   |
| MS6H6     | 12       | 0.35    | 0.48 (0.11 to 2.2) | 30    | 71   |
| ATR       | 11       | 0.049   | 4.7 (1 to 22) | 46    | 54   |
| SLC35F5   | 6        | 0.4     | 2.1 (0.38 to 11) | 17    | 19   |
| MBD4      | 12       | 0.23    | 2 (0.64 to 6.4) | 42    | 59   |
| GRB14     | 10       | 0.32    | 1.9 (0.53 to 6.7) | 38    | 52   |
| CBF2      | 23       | 0.28    | 1.6 (0.69 to 3.7) | 39    | 101  |
| BLM       | 12       | 0.61    | 1.4 (0.38 to 5.2) | 20    | 82   |
| GRK4      | 12       | 0.64    | 1.3 (0.42 to 4.2) | 33    | 61   |
| CDX2      | 10       | 0.97    | 1 (0.26 to 4) | 31    | 66   |
| TCF4      | 12       | 0.36    | 0.54 (0.14 to 2) | 41    | 61   |

Figure 14. Clinical relevance of MSI-driven coding region mutations in target genes in CRC patients. The association of 5 negatively selected MSI-driven mutational events with RFS was calculated in the cohort of 164 MSI CRC patients with survival data available. The association of the Boolean mutational index (see the Materials and Methods section) was calculated from the mutational status of the above 5 target genes (WNK1, PRRC2C, HMGXB4, GART, RFC3) in each tumor sample (status 0, no mutation observed; status 1, at least 1 mutation observed) with RFS also is shown. Also reported is the association with RFS of a series of 15 other frequent MSI mutations (6 positively selected and 9 background events) that we investigated previously in the same MSI CRC samples and published.16 Forest plot of RFS HRs of independent univariate Cox analyses are shown. Squares represent the HRs and horizontal bars represent the 95% CIs. Red indicates a P value of less than 5% (worse prognosis) and blue indicates more than 5%.
(45% of cases), *MSH2* (45%), *MSH6* (~10%), or *PMS2* (~1%). Although MLH1- and MSH2-deficient MSI tumors show similar levels of nucleotide instability, including overall MSI as confirmed here, significantly lower mutation frequencies of R and NR sequences are observed in *MSH6*- and PMS2-deficient MSI tumors. The present cohort was designed to investigate the most common MMR-deficient genotypes in CRC (ie, MLH1- and MSH2-deficient tumors). Only 1 *MSH6*-deficient CRC was included, as mentioned. Future studies could aim to analyze genomic instability in the rare MSI CRC showing *MSH6* or PMS2 deficiency that could show a lower mutation burden at both R and NR sequences.

There is no contradiction between the findings of this study and the literature in the field. Recent data obtained from analysis of thousands of tumors from different primary sites show that, unlike species evolution, positive selection outweighs the negative selection of somatic mutational events during tumor progression. However, these investigations were concerned with alterations at NR sequences and did not take into account mutations in R sequences. In contrast, R sequences (microsatellites) have very high physiopathologic relevance in the tumor model investigated in the present study, namely MSI tumors. We did not observe negative selection of NR sequences in MSI CRC, in line with the prevailing dogma. However, our results show the existence of both positive and negative selection of R sequences during tumor progression, as well as highlighting the dual role for MSI in this important tumor model.

The limitations of our work relate mainly to the analysis of a limited series of MSI CRC using WES, even if it represents a large series of such tumors investigated by this approach. Further studies are required to confirm our results using larger cohorts of MSI CRC, thus allowing the identification of a more robust signature of target genes for MSI that undergo positive or negative selection. Important pro- and anti-cancer genes for MSI tumorigenesis are likely to be included in these genomic signatures. The pathophysiological relevance and opposing functional effects of such MSI-driven events should allow major advances in the understanding of MSI tumorigenesis and in the development of personalized treatments for patients with MMR-deficient tumors.

**References**

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–674.
2. Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481:306–313.
3. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. Universal patterns of selection in cancer and somatic tissues. Cell 2017;171:1029–1041 e21.
4. Bakhoum SF, Landau DA. Cancer evolution: no room for negative selection. Cell 2017;171:987–989.
5. Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Pełtomiaki P, Sistonen P, Aaltonen LA, Nystrom-Lahtii M, Zhang GJ, Meltzer PS, Yu JW, Kao FT, Chen DJ, Cerosaletti KM, Fournier REK, Todd S, Lewis T, Leach RJ, Naylor SL, Weissenbach J, Mecklin JP, Jarvinen H, Petersen GM, Hamilton SR, Green J, Jass J, Watson P, Lynch HT, Trent JM, de la Chapelle A, Kinzler KW, Vogelstein B. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell 1993;75:1215–1225.
6. Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. Science 1993;260:816–819.
7. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 1993;363:558–561.
8. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun 2017;8:15180.
9. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 2016;22:1342–1350.
10. Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 2002;62:2447–2454.
11. Oliveira C, Pinto M, Duval A, Brennetot C, Domingo E, Espin E, Armengol M, Yamamoto H, Hamelin R, Seruca R, Schwartz S Jr. BRAF mutations characterize colon but not gastric cancer with mismatch repair deficiency. Oncogene 2003;22:9192–9196.
12. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 2013;155:858–868.
13. Duval A, Reperant M, Compoint A, Seruca R, Ranzani GN, Iacopetta B, Hamelin R. Target gene mutation profile differs between gastrointestinal and endometrial tumors with mismatch repair deficiency. Cancer Res 2002;62:1609–1612.
14. Woerner SM, Yuan YP, Benner A, Korff S, von Knebel Doeberitz M, Bork P.SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. Nucleic Acids Res 2010;38:D682–D689.
15. Dorard C, de Thonel A, Collura A, Marisa L, Svrcek M, Lagrange A, Jego G, Wanherdrick K, Joly AL, Buhard O, Gobbo J, Penard-Lacronique V, Zouali H, Tubacher E, Kirzin S, Selves J, Milano G, Etienne-Grimaldi MC, Bengrine-Lefevre L, Louvet C, Tournigand C, Lefevre JH, Parc Y, Tiret E, Flejou JF, Gaub MP, Garrido C, Duval A. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. Nat Med 2011;17:1283–1289.
16. Collura A, Lagrange A, Svrcek M, Marisa L, Buhard O, Guilloux A, Wanherdrick K, Dorard C, Taib A, Saget A, Loh M, Soong R, Zeps N, Platell C, Mews A, Iacopetta B, De Thonel A, Seigneuric R, Marcion G, Chapsut C, Lepage C, Bouvier AM, Gaub MP, Milano G, Selves J, Senet P, Delarue P, Arzouk H, Lacoste C, Coquelle A, Bengrine-Lefevre L, Tournigand C, Lefevre JH, Parc Y, Biard DS, Flejou JF, Garrido C, Duval A. Patients with colorectal tumors with microsatellite instability and large...
deletions in HSP110 T17 have improved response to 5-fluorouracil-based chemotherapy. Gastroenterology 2014;146:401–411 e1.

17. Berthenet K, Boudesco C, Collura A, Svrcek M, Richard S, Hammann A, Causse S, Youssi N, Wanherdick K, Duplomb L, Duval A, Garrido C, Jego G. Extracellular HSP110 skews macrophage polarization in colorectal cancer. Oncoimmunology 2016;5:e1170264.

18. Berthenet K, Bokhari A, Lagrange A, Marcion G, Boudesco C, Causse S, De Thonel A, Svrcek M, Goloudina AR, Dumont S, Hammann A, Biard DS, Demidov ON, Seigneuric R, Duval A, Collura A, Jego G, Garrido C. HSP110 promotes colorectal cancer growth through STAT3 activation. Oncogene 2017;36: 2328–2336.

19. Saghier D, Hsu A, Strauss B. Stabilization of the intermediate in frameshift mutation. Mutat Res 1999;423:73–77.

20. Duval A, Rolland S, Compont A, Tubacher E, Iacopetta B, Thomas G, Hamelin R. Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. Hum Mol Genet 2001;10:513–518.

21. Yamamoto H, Imai K. Microsatellite instability: an update. Arch Toxicol 2015;89:899–921.

22. Kondelin J, Gylfe AE, Lundgren S, Tanskanen T, Hamberg J, Aavikko M, Palin K, Ristolainen H, Katainen R, Kaasinen E, Taipale M, Taipale J, Renkonen-Sinisalo L, Jarvinen H, Bohm J, Mecklin JP, Tuupanen S, Aaltonen LA, Pitkanen E. Comprehensive functional analysis of mononucleotide microsatellites in microsatellite-unstable colorectal cancer. Cancer Res 2017;77:4078–4088.

23. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Buhard O, Capel E, Zouali H, Praz F, Muleris M, Renecal K, Rene-Coraill E, Fiaschi D, Ayadi M, Kirzin S, Chazal M, Flejou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Toiky T, Tutt AN, Valdes-Mas R, van Buuren MM, Van’t Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Toiky T, Tutt AN, Valdes-Mas R, van Buuren MM, van’t Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, Mc Dermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Lichter P, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248–5257.

24. Buhard O, Cattaneo F, Wong YF, Yam SF, Friedman E, Flejou JF, Duval A, Hamelin R. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol 2006;24:241–251.

25. Mccullagh P, Nelder JA. Generalized linear models. Boca Raton, FL: CRC Press 1989;37.

26. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–13.

27. Biard DS, Despras E, Sarasin A, Angulo JR. Development of new EBV-based vectors for stable expression of small interfering RNA to mimic human syndromes: application to NER gene silencing. Mol Cancer Res 2005;3:519–629.

28. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248–5257.

29. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–1558.

30. El-Bchiri J, Buhard O, Penard-Lacronique V, Thomas G, Hamelin R, Duval A. Differential nonsense mediated decay of mutated mRNAs in mismatch repair deficient colorectal cancers. Hum Mol Genet 2005;14:2435–2442.

31. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014;30:1015–1016.

32. Williams DA. Extra binomial variation in logistic linear models. Appl Statist 1982;31:144–148.

33. Tibshirani R, Mani J. Robust logistic regression using shift parameters. ALC 2014;2:124–129.

34. Bertsekas DP. Incremental proximal methods for large scale convex optimization. Math Program 2011;129:163.
39. Serysheva E, Mlodzik M, Jenny A. WNKs in Wnt/beta-catenin signaling. Cell Cycle 2014;13:173–174.

40. Cong X, Lu C, Huang X, Yang D, Cui X, Cai J, Lv L, He S, Zhang Y, Ni R. Increased expression of glycinamide ribonucleotide transformylase is associated with a poor prognosis in hepatocellular carcinoma, and it promotes liver cancer cell proliferation. Hum Pathol 2014;45:1370–1378.

41. Liu X, Ding Z, Liu Y, Zhang J, Liu F, Wang X, He X, Cui G, Wang D. Glycinamide ribonucleotide formyl transferase is frequently overexpressed in glioma and critically regulates the proliferation of glioma cells. Pathol Res Pract 2014;210:256–263.

42. He ZY, Wu SG, Peng F, Zhang Q, Luo Y, Chen M, Bao Y. Up-Regulation of RFC3 promotes triple negative breast cancer metastasis and is associated with poor prognosis via EMT. Transl Oncol 2017;10:1–9.

43. Shen H, Cai M, Zhao S, Wang H, Li M, Yao S, Jiang N. Overexpression of RFC3 is correlated with ovarian tumor development and poor prognosis. Tumour Biol 2014;35:10259–10266.

44. de Miguel FJ, Sharma RD, Pajares MJ, Montuenga LM, Rubio A, Pio R. Identification of alternative splicing events regulated by the oncogenic factor SRSF1 in lung cancer. Cancer Res 2014;74:1105–1115.

45. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J, Superti-Furga G, Brummelkamp TR. Gene essentiality and synthetic lethality in haploid human cells. Science 2015;350:1092–1096.