SUPPLEMENTARY METHODS

ASSESSMENT OF THE ANALYTICAL SENSITIVITY OF SNV AND MSI DETECTION BY USING A MOLECULAR BARCODING BASED NGS APPROACH

Cell lines

A HCT15 cell line was cultured in RPMI 1640 medium, whereas RKO and SW480 lines were cultured in DMEM, both supplemented with 10% Fetal Bovine Serum (Thermo Fisher Scientific, USA) and 1% PenStrep (100 U/mL penicillin and 100 µg/mL streptomycin; Thermo Fisher Scientific), maintained in humidified 37°C 5% CO_2 incubators. Genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA) according to manufacturer´s instructions. DNA quality was assessed using a Nanondrop ND 1000 Spectrophotometer (Thermo Fisher Scientific) and agarose gel electrophoresis. Extracted DNA was quantified by Qubit Flurometer with the dsDNA BR Assay (Invitrogen, Carlsbad, CA, USA).

ClearSeq Cancer HS panel

The ClearSeq Cancer HS panel (Agilent Technologies, USA) was used to assess the analytical sensitivity of single nucleotide variants (SNV) and microsatellite instability (MSI) detection. It is a 79.5 kbp commercial panel targeting COSMIC hotspots in 47 genes. The panel utilizes HaloPlexHS, an amplicon-based target enrichment method designed to detect low-frequency allelic variants by the addition of a degenerate 10-nucleotide-long molecular barcode index to the captured DNA fragments.

Sequencing libraries were prepared by using serial dilutions (100%, 50%, 10%, 2%, 0.4%, 0%) of DNA extracted from the microsatellite unstable cell lines (RKO or HCT15) in DNA extracted from the microsatellite-stable SW480 cell line. Library preparation of the target regions in the cell line DNA mixtures was performed using HaloPlex HS Target Enrichment kit (Agilent Technologies, USA), according to the HaloPlex HS Target Enrichment System For Illumina Sequencing Protocol (Version C0, December 2015). Library pools were prepared with four samples for a final concentration of 4 nM. Pools were quantified by TapeStation (Agilent Technologies, USA), denatured, and sequenced on the Illumina MiSeq sequencer using the 150 bp paired-end sequencing protocol using 15 pM seeding concentration and 10% PhiX spike-in with v3 cartridges and 10 pM seeding concentration and 10% PhiX spike-in with v2 cartridges.

Bioinformatic pipeline and estimated limit of detection

Alignment (SureCall pipeline): Pair-ended sequence reads in FASTQ files from Illumina MiSeq were initially processed with Agilent software Surecall v3.5. Afterward, the reads were aligned to the hg19 human reference genome (February 2009 assembly) using the default parameters of bwa-mem. SureCall performs the molecular barcode analysis and the deduplication process.

SNV Calling (SureCall pipeline): The SNPPET algorithm uses sorted BAM files to perform the SNV calling. Variants were filtered for quality differences between alternate and reference alleles, proximity to 3’ ends, and other quality filters. The parameters were set to detect low frequency variants while minimizing false positives.

SNV and microsatellite (MS) indel calling (custom pipeline): Consensus reads generated from 1 to 3 original reads per molecular barcode were filtered out from the sorted BAM file. Sample-level, fully local indel realignment was performed using the Genome Analysis Toolkit- 3.7.0 (GATK). In a first step we created the intervals to be realigned with the "RealignerTargetCreator" tool using as input the sample BAM file, the reference genome and a gold standard for known indels (Mills_and_1000G_gold_standard.indels.b37.sites.vcf). In a second step the reads are realigned over the intervals created in the previous step using the "IndelRealigner tool". In both cases, the default parameters were used. After filtering and realigning, SAMtools version 1.3.1 "mpileup" command generated mpileup files for each of the final BAM files.

SNV (substitutions) were called using VarScan2 version 2.4.3 (http://varscan.sourceforge.net)[1] "mpileup2snp" command. We set the minimum variant frequency to 0.001, the minimum read depth to make a call to 30, and the minimum number of reads supporting the variant to 3. The variants identified by VarScan2 were further filtered using the "fpfilter" tool, a false positive filter included in the variant caller software, for different quality and position parameters. Only variants in target regions were annotated.

Twelve and 9 unique SNVs respectively, not present in SW480 cell line, were called in HCT15 and RKO cell lines. One out of the 12 SNV in 100% HCT15 was only detected by VarScan2 at a low frequency. The frequencies of identified SNVs decreased proportionally to the factor of dilution (Table S8). The limit of detection was established at 0.004 since possible false variants, only present in one point of the series, were detected at very low frequencies (0.001-0.003) (data not shown). Taking into account the two replicas performed with DNA mixtures from HCT15 and SW480, all variants were detected at a frequency below 0.01 (Table S9).

The MSIseq package in R was used to locate MS (only mononucleotide repeat (MNR) type) greater than 5 bp across all sequences included in the panel. This analysis revealed 499 MNRs of a maximum length of 18 bp mostly located in low-covered regions that were further analyzed for MSI. MNR indel calling was performed following a previously described approach.[2] Absolute frequencies for each allele of the MNR were calculated using a custom R script. Seven unstable or polymorphic microsatellites loci fulfilling previous criteria were identified in DNA mixtures from HCT15 and SW480 DNA cell lines (Figure S8A) that were validated by DNA fragment analysis and fluorescent capillary electrophoresis (Figure S8B). Stutter peaks caused by polymerase slippage during PCR or sequencing errors were observed in electropherograms but they were not always present in the analysis by Haloplex NGS, which demonstrated higher specificity. Although some unstable alleles were detected at a very low frequency (below 0.01), the limit of detection needed to be established independently for each microsatellite.

ASSESSMENT OF MICROSATELLITE INSTABILITY AT HIGH SENSITIVITY (hs-MSI) BY USING A CUSTOM PANEL

Custom NGS panel design and sequencing

Two hundred seventy-seven microsatellite targets (91% MNRs) were included in the custom panel. Selected microsatellite target regions included: 1) MS markers frequently mutated in MSI-H tumors: MNR with increased mutation frequency in MSI-H colorectal cancer, gastric cancer, endometrial cancer and colon cell culture tumors according to SelTar*base* (http://seltarbase.org, release 201307)[3], and additional microsatellites that were present in >15% of MSI-H CRC by whole-exome sequencing[4]; 2) MS markers from MSI diagnostic panels: microsatellite loci from the Bethesda panel and Promega MSI Analysis System version 1.2[5], 59 MNR sites included in the reported MSI-detection panel by Zhao et al.[6], and MT1X (T)20[6]; 3) MS within antigen presentation genes according to SelTar*base*; 4) Additional published MS targets of interest.[7, 8, 9, 10]

The custom panel was designed for the HaloPlex HS target enrichment using SureDesign (Agilent, Santa Clara, v4.0.0.18). Target regions included 50bp flanking regions around the microsatellite. Only probes for "on-target" amplicons that covered the entire microsatellite region with 10bp flanks were considered. Microsatellite target regions not covered by more than one "on-target" amplicon were excluded. Library pools were prepared with 26 samples for a final concentration of 25 nM. Pools were quantified by TapeStation (Agilent Technologies, USA), denatured, and sequenced at high depth (20.000x) on an Illumina HiSeq2500 sequencer using the 150 bp paired-end sequencing protocol, 9.5 pM seeding concentration and 25% PhiX spike-in. After deduplication process, the mean depth (±SD) was 1312 ± 447 reads/marker/sample.

Optimization of the bioinformatics pipeline for microsatellite indel calling

The bioinformatics pipeline described above for microsatellite indel calling was further optimized. After SureCall alignment, all reads that did not cover completely the microsatellite were filtered out from the BAM file using SAMtools prior to indel calling. The PCR and sequencing error and/or basal instability for each MS locus was assessed in 22 healthy control blood DNA samples. The minimum read depth for each microsatellite loci was set to 100 and no minimum of reads supporting indels was established in order to capture all PCR and sequencing errors. Only microsatellites with valid data in at least 5 out of 22 control samples (256 out of 277) were considered. Six MS were additionally excluded because the wildtype allele in controls did not match the reference genome. Mean frequencies plus 3 SD were calculated for each microsatellite allele, including the wildtype, and they were used as reference values for case sample analysis.

In order to increase the sensitivity in MSI assessment, 231 out of 256 MS, truly monomorphic in the baseline (with a mean frequency of wildtype allele above 0.94) were selected. Among them, 186 markers (93% MNR 5-14bp of lenght) were previously reported as frequently mutated in MSI-H tumors.[3, 4]

Indel calling in case samples was performed by using the pipeline described above for the analysis of baseline control samples with minor modifications to add more restrictive

parameters: the minimum frequency was set to 0.004 (since we had previously established this limit of detection for SNV), and the minimum number of reads supporting the indel was set to 3.

MSI classification system

To assess the hs-MSI status in each microsatellite locus, the instability level, corresponding to the sum of the frequencies of all allele lengths different from the wildtype (mutational load method), was calculated as (1 – wildtype allele frequency). This value was compared with the reference value (1 – mean wildtype frequency in baseline). Whenever the instability level exceeded the mean value in baseline plus 3 SD and the highest value among the individual samples of the baseline the MS was considered unstable.

As an alternative, the individual allele method was evaluated. Here the frequencies of each alternative microsatellite allele length (instead of the sum of them) were compared against their respective length frequency in the baseline. If at least one of the individual allele length frequencies in a case sample exceeded the baseline mean frequency plus 3 SD and the highest value among the individual samples of the baseline, then the MS was considered unstable.

For both methods, an hs-MSI score was calculated per case sample, representing the percentage of unstable microsatellites out of the total number of valid microsatellite markers (minimum read depth of 100 per microsatellite locus). A threshold for identifying MSI status (positive/negative) in case samples was set to the mean hs-MSI score of the 15 healthy controls included in the training set plus 3 SD.

Hs-MSI score median was compared between different training set groups using a Wilcoxon Rank Sum Test and according to germline affected gene using a Kruskal-Wallis test followed by a Dunn's multiple pairwise comparisons test (Bonferroni correction). MSI score association with age at blood sampling and age at onset was analyzed using the Spearman's correlation coefficient (rho, rs). All the analyses were performed in R software (http://www.R-project.org).

ANALYSIS OF DINUCLEOTIDE REPEATS

Germline MSI (gMSI) analysis

Germline microsatellite instability (gMSI) analysis was performed as described in Ingham et al. 2013[9] in training cohort CMMRD and 22 baseline control blood samples. PCR amplification of the dinucleotide microsatellites D17S791, D2S123, and D17S250 was performed. PCR products were analyzed on an Applied Biosystems 3130XL Genetic Analyzer using GeneMapper software (Applied Biosystems, Forster City, California, USA). The gMSI ratios were determined by dividing the height of an allele's trailing "stutter" peak (n+1) by the height of the allele's major peak (n). A gMSI ratio threshold was chosen for each microsatellite that ensured a high specificity and sensitivity.[9] On the basis of the sensitivity and specificity data, elevated gMSI ratios were defined as follows: >0.060 for D2S123 (mean + 4 SDs), >0.069 for D17S250 (mean + 4 SDs) and >0.117 for D17S791 (mean + 3 SDs). In CMMRD samples, observed gMSI ratio minus the marker-specific threshold was calculated; positive values represented ratios above the

threshold. If two or more markers were above the threshold, the sample was classified as gMSI positive.

Analysis of dinucleotide repeats from NGS data

Analysis of dinucleotide D2S123 was carried out by read counting in IGV v.2.4.10 platform using hg19 reference genome. The zygosity of each marker was previously determined in the gMSI analysis (see above). The level of instability was calculated as the percentage of the sum of reads of all allele lengths different from the wild-type, divided by the wildtype reads x 100. The instability threshold value for D2S123 was defined as >8.23 (mean + 3 SDs).

CLONALITY TESTING OF BLOOD SAMPLES FROM CMMRD PATIENTS

Presence of lymphoproliferative clones in CMMRD patients' blood samples was evaluated using standard BIOMED-2 assay.[11] Briefly, clonally rearranged immunoglobulin and T-cell receptor genes were assayed by multiplex PCR with fluorescence primer sets for IGH VH-JH, IGH DH-JH, IGK, IGL, TCRB, TCRG, and TCRD rearrangements and size-resolved by capillary electrophoresis.

REFRENCES

1 Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22(3):568-76.

2 Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. *Clin Chem* 2014;60(9):1192-9.

3 Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* 2016;22(11):1342-50.

4 Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, Bahl S, Cao Y, Amin-Mansour A, Yamauchi M, Sukawa Y, Stewart C, Rosenberg M, Mima K, Inamura K, Nosho K, Nowak JA, Lawrence MS, Giovannucci EL, Chan AT, Ng K, Meyerhardt JA, Van Allen EM, Getz G, Gabriel SB, Lander ES, Wu CJ, Fuchs CS, Ogino S, Garraway LA. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep* 2016;17(4):1206.

5 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2(5):401-4.

6 Zhao H, Thienpont B, Yesilyurt BT, Moisse M, Reumers J, Coenegrachts L, Sagaert X, Schrauwen S, Smeets D, Matthijs G, Aerts S, Cools J, Metcalf A, Spurdle A, Amant F, Lambrechts D. Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *Elife* 2014;3:e02725.

7 Coolbaugh-Murphy M, Maleki A, Ramagli L, Frazier M, Lichtiger B, Monckton DG, Siciliano MJ, Brown BW. Estimating mutant microsatellite allele frequencies in somatic cells by small-pool PCR. *Genomics* 2004;84(2):419-30.

8 He S, Zhao Z, Yang Y, O'Connell D, Zhang X, Oh S, Ma B, Lee JH, Zhang T, Varghese B, Yip J, Dolatshahi Pirooz S, Li M, Zhang Y, Li GM, Ellen Martin S, Machida K, Liang C. Truncating mutation in the autophagy gene UVRAG confers oncogenic properties and chemosensitivity in colorectal cancers. *Nat Commun* 2015;6:7839.

9 Ingham D, Diggle CP, Berry I, Bristow CA, Hayward BE, Rahman N, Markham AF, Sheridan EG, Bonthron DT, Carr IM. Simple detection of germline microsatellite instability for

diagnosis of constitutional mismatch repair cancer syndrome. *Hum Mutat* 2013;34(6):847-52.

10 Ripberger E, Linnebacher M, Schwitalle Y, Gebert J, von Knebel Doeberitz M. Identification of an HLA-A0201-restricted CTL epitope generated by a tumor-specific frameshift mutation in a coding microsatellite of the OGT gene. *J Clin Immunol* 2003;23(5):415-23.

11 van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, Garcia-Sanz R, van Krieken JH, Droese J, Gonzalez D, Bastard C, White HE, Spaargaren M, Gonzalez M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* 2003;17(12):2257-317.