Received: 10 June 2020 | Revised: 20 January 2021 | Accepted: 21 January 2021

DOI: 10.1002/gcc.22941

RESEARCH ARTICLE

WILEY

No evidence of EMAST in whole genome sequencing data from 248 colorectal cancers

Johanna Kondelin[1,2] | Samantha Martin[1,2] | Riku Katainen[1,2] | Laura Renkonen-Sinisalo[3] | Anna Lepistö[3] | Selja Koskensalo[4] | Jan Böhm[5] | Jukka-Pekka Mecklin[6,7] | Tatiana Cajuso[1,2] | Ulrika A. Hänninen[1,2] | Niko Välimäki[1,2] | Janne Ravantti[1,2] | Kristiina Rajamäki[1,2] | Kimmo Palin[1,2,8] | Lauri A. Aaltonen[1,2,8]

[1]Medicum/Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland

[2]Applied Tumor Genomics Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland

[3]Department of Surgery, Helsinki University Central Hospital, Hospital District of Helsinki and Uusimaa, Helsinki, Finland

[4]The HUCH Gastrointestinal Clinic, Helsinki University Central Hospital, Helsinki, Finland

[5]Department of Pathology, Jyväskylä Central Hospital, Jyväskylä, Finland

[6]Department of Education and Research, Jyväskylä Central Hospital, Jyväskylä, Finland

[7]Department Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

[8]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, Helsinki, Finland

Correspondence
Lauri A. Aaltonen, University of Helsinki, Biomedicum Helsinki, PO Box 63, Helsinki FIN-00014, Finland.
Email: lauri.aaltonen@helsinki.fi

Funding information
Academy of Finland, Grant/Award Numbers: #1312041, #1335823; Biocentrum Finland; HiLIFE; iCAN Digital Precision Cancer Medicine Flagship, Grant/Award Number: 320185; Instrumentarium Research Foundation; Jane and Aatos Erkko Foundation; SYSCOL; The Finnish Cancer Society; The Sigrid Juselius Foundation

Abstract

Microsatellite instability (MSI) is caused by defective DNA mismatch repair (MMR), and manifests as accumulation of small insertions and deletions (indels) in short tandem repeats of the genome. Another form of repeat instability, elevated microsatellite alterations at selected tetranucleotide repeats (EMAST), has been suggested to occur in 50% to 60% of colorectal cancer (CRC), of which approximately one quarter are accounted for by MSI. Unlike for MSI, the criteria for defining EMAST is not consensual. EMAST CRCs have been suggested to form a distinct subset of CRCs that has been linked to a higher tumor stage, chronic inflammation, and poor prognosis. EMAST CRCs not exhibiting MSI have been proposed to show instability of di- and trinucleotide repeats in addition to tetranucleotide repeats, but lack instability of mononucleotide repeats. However, previous studies on EMAST have been based on targeted analysis of small sets of marker repeats, often in relatively few samples. To gain insight into tetranucleotide instability on a genome-wide level, we utilized whole genome sequencing data from 227 microsatellite stable (MSS) CRCs, 18 MSI CRCs, 3 *POLE*-mutated CRCs, and their corresponding normal samples. As expected, we observed tetranucleotide instability in all MSI CRCs, accompanied by instability of mono-, di-, and trinucleotide repeats. Among MSS CRCs, some tumors displayed more microsatellite mutations than others as a continuum, and no distinct subset of tumors with the previously proposed molecular characters of EMAST could be observed. Our results suggest that tetranucleotide repeat mutations in non-MSI CRCs represent stochastic mutation events rather than define a distinct CRC subclass.

KEYWORDS
colorectal cancer, EMAST, next generation sequencing

1 | INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer in Western countries with a mortality rate of nearly 50%.[1] In 1992 to 1993, instability of short repeated nucleotide sequences, microsatellites, was discovered in sporadic and hereditary CRC, and was subsequently named microsatellite instability (MSI).[2] MSI was observed in approximately 15% of CRCs; in virtually all CRCs from Lynch syndrome patients (3% of all CRCs) and in a subset of sporadic cases (12% of all CRCs). Shortly thereafter, MSI was linked to a defect in the DNA mismatch repair (MMR) machinery, one of the central mechanisms involved in the recognition and repair of DNA replication errors. MSI was found to arise from biallelic inactivation of an MMR gene.[2,3] An individual with Lynch syndrome inherits a heterozygous germline mutation in an MMR gene (*MLH1*, *MSH2*, *MSH6*, or *PMS2*) and is therefore highly predisposed to MSI CRC.[2] Sporadic MSI CRCs most often result from biallelic hypermethylation of *MLH1*. Inactivation of the MMR system results in the accumulation of a high number of mutations across the genome, mostly small insertions and deletions (indels) in microsatellites of different orders, including tetranucleotide repeats.[4] Subsequently, MSI has also been observed in other cancer types, for example in 10% to 20% of endometrial and gastric cancers.[5]

In 1997, The National Cancer Institute agreed on a panel of five microsatellites, the Bethesda panel, that consists of two mononucleotide and three dinucleotide microsatellites, as the reference for determining MSI in CRC.[6] Tumors with instability in two or more markers were considered to show a high level of MSI (MSI-H). Tumors with instability in only one marker were considered to show a low level of MSI (MSI-L). Since then, however, MSI-L CRCs have been shown to exhibit clinical and molecular features identical to microsatellite stable (MSS) CRCs.[2,7-12]

Subsequently, another form of repeat instability, elevated microsatellite alterations at selected tetranucleotide repeats, or EMAST, has been suggested to occur in approximately 50% to 60% of all CRCs.[13-16] Unlike for MSI, no consensus criteria for determining EMAST have been established, and different panels of markers with different thresholds for calling EMAST have been used.[15,17] EMAST has been proposed to result from a dysfunction of MSH3 (MutS homolog 3), one of the proteins involved in human MMR.[14,15,18,19] However, the connection between MSH3 and EMAST does not appear as straightforward as the well-established relationship between the loss of expression of MLH1, MSH2, MSH6 or PMS2, and MSI.[15]

To our knowledge, EMAST has not been reported to depend on a total loss of MSH3 expression, and loss of expression of MSH3 has been reported not to correlate with the number of unstable tetranucleotide repeats observed, although this observation was based on only five tetranucleotide repeats.[20] Instead, heterogeneous loss of expression of MSH3—resulting from its reversible shift from the nucleus to the cytoplasm in response to interleukin-6 (IL6) and hypoxia—together with its decreased expression, has been proposed to be the mechanism behind EMAST.[15,20,21] However, heterogeneous expression of MSH3 has also been observed in CRCs not showing EMAST.[20] To our understanding, a causal relation between somatic *MSH3* mutations or epimutations and EMAST has not been proven.[15] Compound heterozygous germline mutations in *MSH3* have been reported in two unrelated patients with adenomatous polyposis, resulting in a complete loss of MSH3 as observed by immunohistochemical staining.[22] EMAST was suggested to be present in adenomas of these patients.

Tetranucleotide instability appears to occur in all MSI CRCs, where a defect in *MLH1*, *MSH2*, *MSH6*, or *PMS2* leads to an accumulation of indels in repeats.[14,20,23] Yet, MSI CRCs and EMAST CRCs have been suggested to constitute two separate entities, and among CRCs that do not exhibit MSI, EMAST tumors have been proposed to form a subset of CRCs that show instability of tetra-, di-, and trinucleotide repeats, but not mononucleotide repeats.[15,17,24] In contrast to MSI, EMAST has been suggested to modify tumor behavior rather than participate in the initiation of tumorigenesis.[15,25] In CRC, EMAST has been reported to be linked to a higher tumor stage, chronic inflammation, and poor prognosis.[15,20,26] In addition to CRC, EMAST has been suggested to occur in other cancer types, including lung, ovarian, prostate, renal, endometrial, nonmelanoma skin cancer, and cancers of the head and neck with prevalences varying from 9% to 75%.[17] EMAST has been suggested to relate to exposure to environmental carcinogens.[17] It has also been proposed to show potential for serving as a prognostic or preventive biomarker.[16,17,27]

In the past, MSI and EMAST studies were largely based on PCR and subsequent fragment analysis of selected repeat sites, as there was no technology to enable genome-wide depiction of repeat instability. In the past decade, however, next generation sequencing (NGS) technologies have been widely accepted in both research and clinical use. These technologies finally enable the genome-wide characterization of mutations in tumors, including many of the repetitive regions. To date, several large-scale sequencing efforts in CRC have been published.[8-10,12,28-41] These studies have confirmed MSI CRCs as a clearly distinct subset of CRCs, especially in terms of the number of indels in short repeated regions.[8-10,12,28,29,31,33,34,41] However, none of these large-scale studies have had a particular focus on tetranucleotide repeats. In order to comprehensively characterize tetranucleotide repeat instability on a genome-wide level, we utilized whole genome sequencing (WGS) data from 227 MSS CRCs, 18 MSI CRCs, 3 *POLE*-mutated CRCs, and their respective normal samples. To our knowledge, this is the first effort focusing on EMAST at the genomic level. We identified 561 490 tetranucleotide repeats in the human reference genome, and indels of a multiple of four in length were found in 30 306 of them in our WGS data. In our data, however, striking tetranucleotide instability was only observed in MSI CRCs, coincident with instability in mono-, di-, and trinucleotide repeats. Among MSS CRCs, no distinct subgroup of tumors with characteristics fitting EMAST was found. In order to confirm these observations, we performed PCR and subsequent fragment analysis of five tetranucleotide markers in 18 MSI, 3 *POLE*-defective, and 40 MSS CRCs. Also in the fragment analysis data, MSI CRCs formed a clear subset of samples with the most tetranucleotide instability. Some MSS samples showed occasional tetranucleotide mutations, and these tumors tended to be

the ones displaying many short repeat mutations in WGS data in general.

2 | MATERIALS AND METHODS

2.1 | Ethics approval

This study was approved by the Ethics Committee of the Hospital District of Helsinki and Uusimaa, and conducted in accordance with the Declaration of Helsinki. For all samples, signed informed consent was given by the patient, or authorization was received from the National Supervisory Authority for Welfare and Health.

2.2 | Patient material

In this study, fresh frozen specimens of colorectal adenocarcinomas, and corresponding normal colorectal tissue samples or blood from 248 CRC patients in Finland were analyzed. These samples were derived from a population-based series of CRCs from 1042 patients, and a second series of CRCs from two regional hospitals for which collection is ongoing.[42]

The MSI status of the tumors had been previously determined by radioactive labeling techniques, fluorescence-based PCR methods or fragment analysis in previous studies.[42-44] In the radioactive labeling techniques, seven markers (D5S404, D17S787, D5S346, D1S216, D11S904, D10S197, and TP53) were analyzed by two reviewers. A sample was called MSI if 2/7 markers were unstable. If none of the markers were unstable, the sample was called MSS given that at least 5/7 markers were successfully analyzed. If 1/7 markers were unstable, more markers (DCC, D13S175, D7S519, D20S100, D15S120, D2S136, and D14S79) were analyzed so that at least 10 markers were reviewed in total. If one or more of the extra markers were unstable, the sample was called MSI. If none of the markers were unstable, the sample was called MSS.

In cases where a fluorescence-based PCR method was utilized, 16 markers (D8S254, MYC, NM23, D5S346, TP53, D1S228, D8S261, D7S496, D8S137, DCC, D7S501, MCC, D5S318, D1S507, D19S394, and RB1) were tested for. If at minimum 30% of the alleles were unstable, the sample was called MSI. Later, two markers (BAT26 and TGFBRII) were utilized. Both markers were evaluated by two independent reviewers. If BAT26 showed deletions, the result was compared to that of the normal sample to ensure the change was of somatic origin.

When fragment analysis was utilized, the Bethesda panel of five markers (BAT25, BAT26, D5S346, D17S250, and D2S123) was analysed.[6] If at least 2/5 markers showed instability, the sample was called MSI.

Of the 248 tumors, 18 fulfilled the criteria for MSI-H, and 230 were MSS, of which 2 were from patients with ulcerative colitis, and three displayed an ultramutator phenotype caused by somatic *POLE* defects (Supplementary Table 1).[29] All MSI samples were sporadic; no germline mutation in an MMR gene had been detected in these patients, the patients typically had no family history of CRC, and did not develop CRC at a young age (Supplementary Table 1). These tumors are known to nearly always relate to biallelic somatic hypermethylation of the *MLH1* promoter.[45] Detailed clinical information and a pathologist's evaluation were available for all samples.

2.3 | Whole genome sequencing

DNA was extracted from either fresh frozen tissue or blood using standard methods. Whole-genome sequencing was carried out for the CRC samples, and their corresponding normal pairs.[46] Paired-end sequencing was performed with Illumina HiSeq 2000 as an Illumina service, or Illumina HighSeq X Ten as a SciLifeLab service. Read lengths were 100 and 151 bp, and the median coverage was 47.6 and 28.3, respectively.

2.4 | Somatic variant calling and quality control

Primary analysis and somatic variant calling were performed with GATK4 best practices workflow (version 4.0.4.0.) for all tumor/normal pairs. The GRCh38 reference genome was used in all analyses. All variants that were given a "PASS" filter value by Mutect2 were included. In order to include all somatic variants at repeat sites, variants annotated with "str_contraction" and "panel_of_normals" filter values were also included.

2.5 | Variants in different repeat regions

We included all tetranucleotide repeats in the genome with at least three consecutive repeat units. For the tetranucleotide repeat analysis, all other repeat orders within tetranucleotide repeats were excluded (eg, mono- and dinucleotide repeats). We considered all somatic indel calls that were located precisely at the start of the repeat region and that were multiples of four-base pairs in length (ie, the length of a tetranucleotide repeat unit). For a subsequent analysis, mono-, di-, and trinucleotide repeats were studied similarly, with the exception of mononucleotide repeats where at least five consecutive repeat units were required. BasePlayer was used for the annotation and visual validation of repeats and variants.[47]

2.6 | Data plotting

Data was plotted with R using the ggplot2 library.[48,49] The maximum likelihood negative binomial fit was performed with the function glm.nb from the MASS package (Mean = 63.38, dispersion = 4.79).[50]

2.7 | Fragment analysis of tetranucleotide markers

Fragment analysis was performed for 60 tumor-normal pairs; 17 MSI, 3 POLE-defective, and a subset of 40 MSS samples. Fragment analysis

TABLE 1 Primer sequences of the EMAST microsatellite markers

Marker	Forward primer	Reverse primer
MYCL1	TGGCGAGACTCCATCAAAG	CCTTTTAAGCTGCAACAATTTC
D20S85	GAGTATCCAGAGAGCTATTA	ATTACAGTGTGAGACCCTG
D20S82	GCCTTGATCACACCACTACA	GTGGTCACTAAAGTTTCTGCT
D8S321	GATGAAAGAATGATAGATTACAG	ATCTTCTCATGCCATATCTGC
D9S242	GTGAGAGTTCCTTCTGGC	ACTCCAGTACAAGACTCTG

was also performed for one unpaired MSI tumor where normal tissue was no longer available. The MSS tumors were uniformly selected based on the number of unstable tetranucleotide repeats identified from the WGS data: samples were ranked from highest to lowest by the number of indels in their tetranucleotide repeats and approximately every sixth sample was selected for fragment analysis, based on availability of sample material. Five tetranucleotide markers—MYCL1, D20S85, D20S82, D8S321, and D9S242—were amplified by PCR (Table 1).[51] Each PCR reaction contained 0.15 μL AmpliTaqGold (Applied Biosystems, Waltham, MA), 1.5 μL buffer (Applied Biosystems), 0.2 μL dNTPs (BioNordika, Helsinki, Finland), 0.6 μL fluorescent-tagged forward primer (ThermoFisher Scientific, Life Technologies, Waltham, MA), 0.6 μL reverse primer (Sigma-Aldrich, St. Louis, MO), 9.95 μL water, and approximately 10 ng DNA extracted from a fresh-frozen tumor. The PCR conditions consisted of an initial denaturation at 95°C for 10 minutes, followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing at 60°C for 75 seconds, and extension at 72°C for 1 minute, before a final extension step at 72°C for 30 minutes.

Fragment analysis was performed by capillary electrophoresis at the Institute for Molecular Medicine Finland Technology Centre (FIMM; Helsinki, Finland) with the ABI3730XL DNA Analyzer with GeneScan 500 LIZ size standard (Applied Biosystems). The manufacturer's instructions were followed for all methods.

GeneMarker software (SoftGenetics, State College, PA; Version 1.40) was used for analyzing the sequencing graphs. Tumor samples were compared to their corresponding normal sample. A marker was considered unstable if a fragment length difference between the tumor and normal sample of a multiple of 4 bp was seen. Fragment analysis peaks in the tumor DNA were called if they had a height of at least 20% that of the adjacent wildtype allele peak for 4 bp deletions, and at least 10% for 4 bp insertions (Supplementary Figure 1). These cut offs were not used for indels of 8 bp or longer as we did not observe stutter peaks of these lengths in the normal DNA. A sample was considered to show tetranucleotide instability when two or more markers were unstable.

3 | RESULTS

In order to comprehensively characterize tetranucleotide repeat instability on a genome-wide level, we utilized WGS data from 227 MSS CRCs, 18 MSI CRCs, 3 *POLE*-mutated CRCs, and their respective normal samples. Indels in tetranucleotide repeats were evaluated to determine whether any tumor subgroups with EMAST could be found. In total, 561 490 tetranucleotide repeats were identified in the reference genome. Across all 248 CRCs, a total of 49 040 indels of a multiple of four bases in length were observed in 30 306 different tetranucleotide repeats.

3.1 | MSI CRCs harbored the most indels in tetranucleotide repeats

The number of somatic indels in tetranucleotide repeats was compared across all CRCs. MSI CRCs were clearly distinct from MSS CRCs, with MSI tumors consistently containing a higher number of indels (Figure 1A). The median number of indels in tetranucleotide repeats in MSI tumors was 1866, while only 58 in MSS tumors. The number of tetranucleotide indels in MSI CRCs was more variable with over dispersion approximately 5-fold that of the MSS CRCs (95% CI bootstrap estimates [7.7-13.5] and [25.5-123] for MSS and MSI, respectively).

3.2 | No subset of MSS CRCs with tetranucleotide instability was observed

Among MSS CRCs, distinct groups of tumors with differing numbers of indels in tetranucleotide repeats were not found. Instead, the number of indels observed in tetranucleotide repeats was consistent with random sampling from a negative binomial distribution (Kolmogorov-Smirnov test, P-value > .05; Figure 1B). Similarly, when considering the proportion of indels found in tetranucleotide repeats compared to all indels in MSS CRCs, there was continuous variation across tumors and a small number of samples with a very slightly higher or lower proportion than expected by binomial distribution (Supplementary Figure 2).

3.3 | *POLE*-mutated CRCs harbored a large number of indels in tetranucleotide repeats

Three MSS tumors were *POLE*-mutated and had a tetranucleotide repeat indel count in between that of the MSI and the remaining MSS tumors (Figure 1A). They show greater similarity to other MSS tumors in this regard. The proportion of indels in tetranucleotide repeats in these samples, however, was particularly low in comparison to other

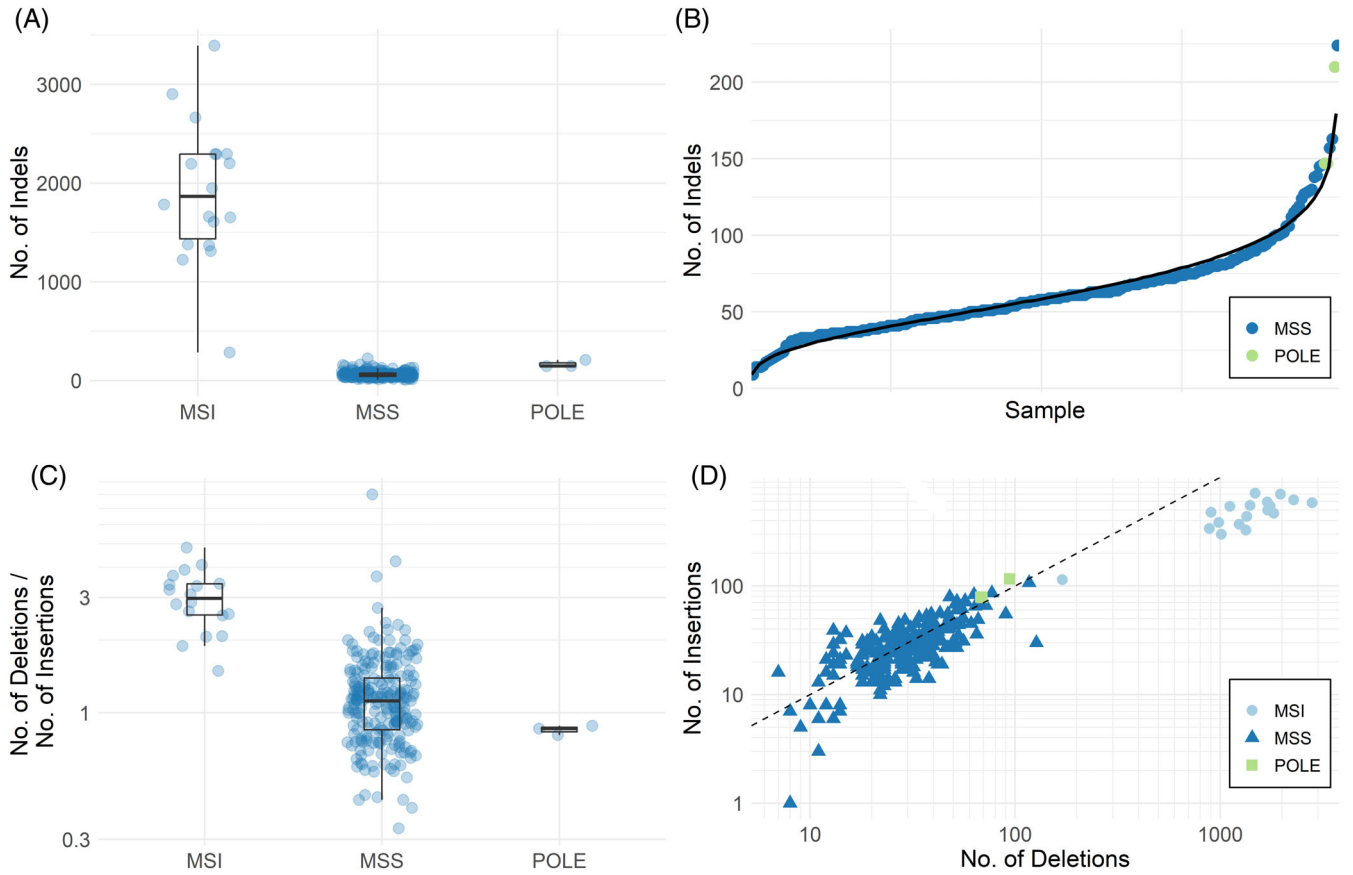

FIGURE 1 Somatic indels in tetranucleotide repeats. A, The number of all somatic indels in tetranucleotide repeats in all 248 CRCs. B, Q-Q plot showing the observed number of somatic indels in tetranucleotide repeats in 227 MSS CRCs and three *POLE*-mutated CRCs. The solid black line represents the expected distribution. C, The number of somatic deletions compared to the number of somatic insertions in tetranucleotide repeats in all 248 CRCs. The Y-axis is on a log scale. D, A log-log plot showing the number of somatic insertions and somatic deletions in all 248 CRCs. The dashed line represents a 1:1 ratio of insertions and deletions. CRC, colorectal cancer; MSS, microsatellite stable

MSS CRCs, and was instead indistinguishable from the MSI CRCs (Supplementary Figure 3). Because both germline and somatic *POLE*-mutations have been shown to coexist with somatic MMR gene mutations and result in MSI CRC, we looked at somatic mutations in MMR genes in these three samples (Supplementary Table 2).[9,52-54] Each sample contained subclonal somatic mutations in several MMR genes, and at least one nonsense mutation in an MMR gene was found in each sample. These subclonal changes provide a conceivable explanation for the observed rate of repeat instability.

3.4 | Only a small proportion of indels were located in tetranucleotide repeats, and deletions were more prominent than insertions

In order to take into account the distribution of indels throughout the genome, the proportion of somatic indels located within tetranucleotide repeats was calculated for each tumor. With the exception of the *POLE*-mutated MSS samples, the tumors still showed clear separation based on their MMR status (Supplementary Figure 3). Although the number of indels in tetranucleotide repeats was higher in MSI tumors, they made up a smaller proportion of the indels throughout the genome. A median of 0.36% and 1.14%, and a maximum of 0.45% and 2.65% of indels were located in tetranucleotide repeats in MSI CRCs and MSS CRCs respectively. Therefore, in all tumors, only a small minority of indels were located in tetranucleotide repeats.

The majority of tumors harbored more deletions than insertions within tetranucleotide repeats (Figures 1C,D). The difference was particularly prominent in MSI CRCs in which a median of only 25.2% of indels were insertions (17.2%-40.1%). In MSS CRCs, the difference was more subtle with a median of 47.2% of indels being insertions (11.1%-75%). Approximately a third of MSS CRCs (87/230) contained a higher number of insertions than deletions (Figures 1C,D).

3.5 | Longer tetranucleotide repeats and simple tetra motifs were the most prone to indels

A large number of indels are located in the shortest tetranucleotide repeats (Supplementary Figure 4). However, when the number of tetranucleotide repeats of each length in the reference genome is

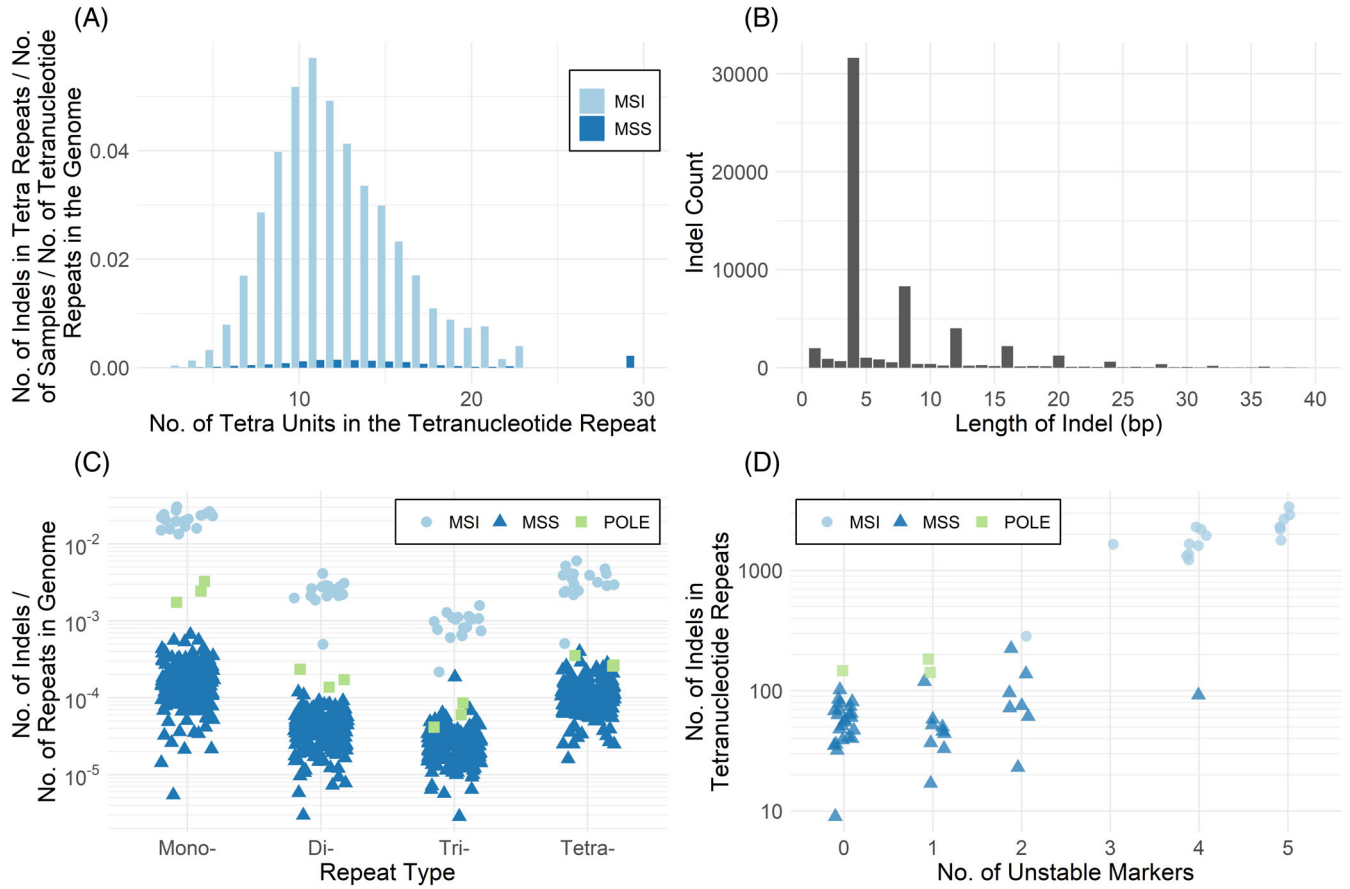

FIGURE 2 Somatic indels in mono-, di-, tri-, and tetranucleotide repeats. A, The number of somatic indels in tetranucleotide repeats of different lengths in all 248 CRCs as a proportion of the number of tetranucleotide repeats in the GRCh38 reference genome. MSI samples—the number of indels in tetranucleotide repeats in the 18 MSI CRCs/18/the number of such tetranucleotide repeats in the reference genome. MSS samples—the number of indels in tetranucleotide repeats in the 227 MSS CRCs and three *POLE*-mutated CRCs/230/the number of such tetranucleotide repeats in the reference genome. B, The lengths of the somatic indels observed in tetranucleotide repeats in all the 248 CRCs up to 40 bp. Unlike elsewhere in this study, in this figure, indels of all lengths are considered instead of only indels that are a multiple of four bases in length. C, The number of somatic indels located within mono-, di-, tri-, and tetranucleotide repeats in all 248 CRCs as a proportion of the total number of corresponding repeats in the GRCh38 reference genome. Mononucleotide repeats of at least 5 units in length were included, and di-, tri-, and tetranucleotide repeats of at least 3 units in length. The Y-axis is on a log scale. D, The number of tetranucleotide repeat indels identified in WGS data in relation to the number of unstable tetranucleotide markers observed in fragment analysis for 17 MSI, 3 *POLE*-defective, and 40 MSS tumors. The Y-axis is on a log scale. CRC, colorectal cancer; MSI, microsatellite instability; MSS, microsatellite stable

taken into account, only a small proportion of the short tetranucleotide repeats are targeted by indels (Figure 2A). Instead, when the number of tetranucleotide repeats of each length in the reference genome is accounted for, tetranucleotide repeats consisting of 11 tetra units are most frequently targeted by indels. As the number of repeated units increases beyond 11, a sharp drop in indel counts is observed. This distribution is seen in both MSS and MSI CRCs (Figure 2A). In longer repeats, indels are likely to be underrepresented due to limitations in current sequencing technologies.

A high proportion of tetranucleotides of 29 repeats appear to contain indels (Figure 2A); however, this is due to the small number of repeats in the genome of this length. Only two were identified in the reference genome, one of which contained an indel in a single tumor. Tetranucleotide repeats of particular motifs contained a higher number of indels than others (Supplementary Table 3). Repeats of simple motifs such as AAAT, TTTA, AAAG, and TTTC contained the highest number of indels.

3.6 | Short indels were more common in tetranucleotide repeats

In this study, only indels that are a multiple of four nucleotides in length were considered in the analysis. Figure 2B depicts indels of all lengths that were observed in tetranucleotide repeats in the 248 CRCs. This figure illustrates that the vast majority of indels in tetranucleotides are indeed a multiple of four nucleotides in length, with their frequency decreasing as the indel length becomes longer. Insertions in particular are preferentially shorter in length (Supplementary Figure 5). Four-base deletions were the most common indels observed.

TABLE 2 Somatic indels in tetranucleotide repeats in protein coding genes in 248 CRCs

Gene	Sample	MSI/MSS/POLE	Position (GRCh38)	Base change	Genotype (calls/coverage)	Allelic fraction	Description
MEPE	s26.1 T1	MSI	4:87845066	G- > del4	Het(20/69)	0.29	Matrix extracellular phosphoglycoprotein (Source: HGNC Symbol%3BAcc: HGNC:13361)
BRCA2	c206.1 T1	MSS	13:32332778	A- > del4	Het(11/50)	0.22	BRCA2%2C DNA repair associated (Source: HGNC Symbol%3BAcc: HGNC:1101)
ZNF66	c39.1 T	MSI	19:20799053	T- > del4	Het(4/20)	0.20	Zinc finger protein 66 (Source: HGNC Symbol%3BAcc:HGNC:13135)
DLG3	c777.1 T	MSI	X:70492526	A- > del4	Het(4/37)	0.11	Disks large MAGUK scaffold protein 3 (Source: HGNC Symbol%3BAcc: HGNC:2902)

Abbreviations: MSI, microsatellite instability; MSS, microsatellite stable.

3.7 | Indels in tetranucleotide repeats in four protein coding genes were observed

The majority of tetranucleotide repeats are in noncoding regions of the genome. However, 207 of the tetranucleotide repeats that were identified overlap with the coding sequence of protein coding genes (Supplementary Table 4). A deletion was observed in four genes: *BRCA2*, *DLG3*, *MEPE*, and *ZNF66* (Table 2). Each gene contained a deletion in only one tumor.

3.8 | Instability of mono-, di-, and trinucleotide repeats was observed in MSI CRCs but not in MSS CRCs

Because EMAST CRCs have been suggested to show instability in di- and trinucleotide repeats in addition to tetranucleotide repeats, but absence of instability in mononucleotide repeats, we looked at the number of indels observed in mono-, di-, and trinucleotide repeats in our 248 CRCs (Supplementary Table 5).[15,24] All MSI CRCs showed a high indel count in mono-, di-, and trinucleotide repeats in addition to tetranucleotide repeats (Figure 2C). In MSI CRCs, a particularly high proportion of indels were observed in mononucleotides, a median of 91.4%, as opposed to 4.6%, and 0.2% in di-, and trinucleotides, respectively. In MSS CRCs, no subsets of tumors were observed to have a higher indel count for any of these repeat lengths (Figure 2C). In MSS CRCs, a lower proportion of indels were observed in mononucleotide repeats than in MSI CRCs (median 69.5%), whereas a higher proportion of indels was observed in di- and trinucleotide repeats (median 7.6% and 0.7%, respectively; Supplementary Figures 6-8). When the number of indels in mono-, di, tri-, and tetranucleotide repeats was compared to the number of such repeats in the genome, mononucleotide repeats are the most mutated, followed by tetranucleotide, dinucleotide, and trinucleotide repeats (Figure 2C, Supplementary Table 6).

A simple linear regression was applied to the MSS samples and the number of tetranucleotide repeat indels was found to be strongly correlated with both the number of mononucleotide (adjusted R^2 0.6213, $P < 2e^{-16}$) and dinucleotide repeat indels (adjusted R^2 0.5251, $P < 2e^{-16}$; Supplementary Figure 9).

3.9 | MSS tumors with a high number of indels tend to display a proximal location, higher age, and higher sequencing coverage

Clinical characteristics were compared between the MSS tumors in the top and bottom thirds by the number of mono-, di-, tri-, and tetranucleotide repeat indels (Supplementary Table 7). No significant difference was observed at a significance level below 0.05 in regards to sex, Duke's stage, tumor grade, or the immune cell score (T cell infiltration) with any repeat unit length. The tumor location (Fisher's exact test, all $P < .027$), age of the patient at diagnosis and the average read coverage on variant in WGS data (Wilcoxon rank sum test, all $P < .022$ and $P < .021$, respectively) were identified as being significantly different with all repeat unit lengths. MSS tumors in the upper third of the spectrum more often tended toward a proximal location (Supplementary Figure 10), higher age (for tetra indels, median 71 and 69, and range 50-91 and 28-87, for top and bottom thirds, respectively), and a higher average coverage (for tetra indels, mean 45.7 and 42.5, and SD 9.21 and 9.57, for top and bottom thirds, respectively) than tumors with a high number of indels.

3.10 | The number of unstable tetranucleotide markers in fragment analysis does not fully reflect genome-wide tetranucleotide instability in WGS data

Fragment analysis was performed for all 18 MSI and 3 *POLE*-defective, and 40 of the MSS tumors selected to represent samples across the whole range of tetranucleotide repeat indel counts (Figure 2D). Five tetranucleotide markers (*MYCL1*, D20S85, D20S82, D8S321, and D9S242) traditionally used to call EMAST were utilized.[51] The unstable markers for each tumor are shown in Supplementary Table 8.

Consistent with the WGS data, all 18 MSI tumors showed tetranucleotide instability in fragment analysis, defined here as instability in two or more of the five markers (Figure 2D). For one of the MSI tumors, matching normal DNA was not available for fragment analysis, but the tumor harbored three alleles for two of the markers in the tumor DNA and so was inferred as unstable. The majority of MSI tumors showed instability in four or five markers and only two in fewer than four markers (Figure 2D). The majority of MSS samples and all three *POLE*-defective samples showed instability in zero or one markers (57.5% and 22.5%, respectively, for MSS samples). Thus, consistent with the WGS data, these samples did not show tetranucleotide instability according to fragment analysis. However, seven of the MSS tumors showed instability in two, and one MSS tumor in four markers, and thus, in contrast to WGS, appeared to show tetranucleotide instability in fragment analysis (Figure 2D). Overall, in the fragment analysis, 20% of MSS tumors showed tetranucleotide instability while in the WGS data no tetranucleotide instability was observed in these tumors.

The MSS tumors with two or four unstable markers in fragment analysis were typically in the upper end of the spectrum of WGS tetranucleotide repeat indel counts, while tumors with only one unstable marker were typically in the lower end (Supplementary Figure 11). Notably, MSS tumors with no unstable markers were interspersed among the MSS tumors with unstable markers throughout the whole spectrum of WGS tetranucleotide repeat indel counts (Supplementary Figure 11).

In MSS samples, each additional unstable marker associates with approximately 20% more (10%-34%, Bayesian Negative Binomial model, 95% High-Density Interval) tetranucleotide repeat indels in WGS on average and the distributions of indel counts with a given number of unstable markers are highly overlapping (Figure 2D). None of the five markers are significantly correlated with any (Fisher's exact test) or all (logistic regression) others in our dataset of MSS samples.

4 | DISCUSSION

Previously, the study of repetitive regions of the genome mostly consisted of targeted fragment analysis of selected repeat sites as no technology enabling genome-wide characterization of repeat mutations was available. MSI was discovered in the early 1990s through shortening or lengthening of short tandem repeats in approximately 15% of CRCs, and was soon linked to defective DNA MMR[2]. Subsequently, another form of repeat instability, EMAST, has been suggested to occur in approximately 50% to 60% of which approximately one quarter are accounted for by MSI CRCs.[13-16,20,23] The emergence of NGS and the ensuing large-scale sequencing studies have confirmed mutation accumulation in the genome to be non-uniform, and the mutability of repeated regions is affected by factors such as their length, nucleotide composition and genomic location.[9,10,12,30,31,55] Therefore, genome-wide studies are required to obtain a comprehensive picture of repeat instability in cancer. Large-scale sequencing studies on CRC have confirmed MSI CRCs to be a distinct subset of CRC with a large number of mutations, especially indels in their short tandem repeats.[8-10,12,28,29,31,33,34,41] To our knowledge, however, EMAST has not been studied on the genome-wide level before. Hence, in order to gain a comprehensive picture of tetranucleotide repeat instability in CRC, we utilized WGS data from 227 MSS CRCs, 18 MSI CRCs, 3 *POLE*-mutated CRCs, and their respective normal samples.

We identified 561 490 tetranucleotide repeats in the human reference genome and from these found 30 306 indels of a multiple of four bases in length in our WGS data. In accordance with previous large-scale sequencing efforts in CRC, MSI was evident in our WGS data as the MSI tumors harbored the most indels in repeat regions (Supplementary Table 3).[8-10,12,28,29,31,33,34,41,43,44] Instability of tetranucleotide repeats was observed in all 18 MSI CRCs as anticipated (Figure 1A).[14,20,23] Indels in CRCs exhibiting MSI have been reported to occur most prominently in mononucleotide repeats, but also in other short repeats. This was observed in our data as the MSI tumors exhibited indels in mono-, di-, and trinucleotide repeats in addition to tetranucleotide repeats (Figure 2C).[9,15,24,31] When the number of indels observed was compared to the number of such repeats in the reference genome, mononucleotide repeats were the most highly mutated, followed by tetranucleotide, dinucleotide, and trinucleotide repeats (Figure 2C). This is most likely explained by a relatively high number of longer, more unstable, tetranucleotide repeats in the human genome.

It has been suggested that a group of CRCs that does not exhibit MSI shows EMAST.[13-16] However, in our WGS data, no subset of MSS CRCs with specific tetranucleotide instability was identified (Figures 1A,B). Overall, in MSS CRCs, tetranucleotide repeats were fairly stable with a median of 58 indels in tetranucleotide repeats per tumor, compared to a median of 1866 in MSI CRCs. Of note, the sequencing technology used is not suitable for observing very long tetranucleotide repeats, such as those that have traditionally been used for determining EMAST (*MYCL1*, D20S85, D8S321, D20S82, and D9S242), because their length hampers alignment of the sequence reads.[26,31] In order to confirm this observation, we performed PCR and subsequent fragment analysis of these five tetranucleotide markers in 18 MSI, 3 *POLE*-defective, and 40 MSS CRCs. Also in this data, only the MSI CRCs formed a clear subset of samples showing the highest numbers of unstable markers (typically four or five) and consistent tetranucleotide instability (Figure 2D). Among MSS CRCs, no distinct subset of tumors with EMAST could be detected, the majority showing no unstable markers while several tumors showed one or two (Figure 2D). In MSS samples, each additional unstable marker associates with proportional increase in the number of tetranucleotide repeat indels in WGS, and instabilities in the five markers appear uncorrelated. Taken together, all this indicates the marker instability in MSS tumors as a proxy for the continuum of tetranucleotide repeat indel counts, instead of being discriminant of a separate group of tumors. Furthermore, this is consistent with the Negative Binomial distribution of the indel counts in population (Figure 1B) being formed by Gamma-Poisson mixture, where each repeat mutation appears independently from others, given a tumor

specific, Gamma distributed, base mutation rate. The variation of this base mutation rate can be largely attributed to features such as age of the patient, tumor location, and sequencing depth. A higher number of tetranucleotide repeat indels in MSS tumors WGS data was also a predictor for a higher number of mono- and dinucleotide repeat indels in the same tumor. Therefore, based on our results, the presence of one or two unstable tetranucleotide markers in fragment analysis—both criteria traditionally used to detect EMAST—simply predicts a somewhat higher number of short repeat mutations in general.[15,17]

EMAST CRCs have been suggested to show instability in di-, and trinucleotide repeats in addition to instability in tetranucleotide repeats, but absence of mononucleotide repeat instability.[15,24] In our WGS data, however, striking tetranucleotide instability was always observed together with instability in mono-, di-, and trinucleotide repeats, and solely in MSI CRCs (Figure 2C). Our observation that tetranucleotide instability is observed in MSI CRCs is compatible with two other large-scale CRC sequencing studies where indels in tetranucleotide repeats were observed.[9,31] Instability of dinucleotide repeats with no instability in mononucleotide repeats has been claimed to be a hallmark of the so-called MSI-L phenotype, and MSI-L and EMAST have been suggested to be either the same or an overlapping phenomenon.[14,15,23,56] In our WGS data, however, no such subset of tumors with sole dinucleotide instability was observed (Figure 2C). This, too, is in line with previous large-scale CRC sequencing efforts.[9,10] The number of tetranucleotide repeat indels in MSS tumors was found to be strongly correlated with the numbers of mono- and dinucleotide repeat indels, reinforcing the proposed overlap of the phenomena reported previously as EMAST and MSI-L. The findings from our data support the view that repeat mutations in tumors without MSI represent stochastic mutation events with the presence of intact MMR systems, and MSI-L and EMAST as such are not biologically relevant subtypes of CRC.

The clinical correlates between tumors of high and low-repeat indel count were highly similar when analyzing repeats of any unit length, with tetranucleotide repeats in no way distinguishing from mono-, di-, or trinucleotide repeats. The slight excess number of repeat indels in MSS tumors on the upper end of the spectrum may partially be due to the observed higher age of these patients at diagnosis, allowing increased time for mutations to accumulate, as well as the higher average sequencing coverage. Tumors with a high number of repeat indels of any unit length were located proximally slightly more often than tumors with a low-repeat indel count. Although we detected no difference in tumor stage between these groups, this tendency could be related to tumor age, proximal tumors perhaps requiring a longer time from initiation to removal.

In the WGS data, among the MSS CRCs, the three *POLE*-mutated CRCs were amidst those with the highest number of indels in tetranucleotide repeats, and the same was observed for mono-, di-, and trinucleotide repeats (Figures 1A and 2C). Strikingly elevated single nucleotide variation (SNV) rates in particular have been previously reported in *POLE*-mutated tumors.[8] However, in a study by Kim et al, one *POLE*-mutated CRC genome and one *POLE*-mutated endometrial genome were reported to harbor *MLH1* silencing and result in an MSI phenotype.[9] Subsequently, both germline and somatic *POLE*-mutations have been shown to coexist with somatic MMR gene mutations and result in MSI CRC.[52-54] Indeed, in our data, we observed somatic subclonal mutations in MMR genes in the *POLE*-mutated samples (Supplementary Table 2). The resulting MMR deficient subclones provide a plausible explanation as to why *POLE* tumors, despite in general being MSS, display relatively many repeat mutations in the genome-wide NGS data.

According to Watson et al, in most EMAST cancers, instability occurs at loci with AAAG or ATAG motifs.[17] Also in our WGS data, these were among the most highly mutated motifs (Supplementary Table 3). Of the indels observed in tetranucleotide repeats, deletions were more common than insertions, especially in MSI CRCs (Figures 1C,D). In tumors with MSI as well as in the *POLE*-mutated tumors, a smaller proportion of all indels was observed in tetranucleotide repeats than in MSS samples (Figure 2A and Supplementary Figure 3). For the MSI CRCs, this was not surprising given that mononucleotide repeats have been shown to accumulate the most indels in these tumors.[9,31]

We identified 207 tetranucleotide repeats in protein coding genes in the human reference genome (Supplementary Table 4). Of these, indels were observed in only four genes (*BRCA2*, *MMPE*, *DLG3*, and *ZNF66*), of which all contained a deletion in one tumor (Table 2). *BRCA2* (*BRCA2 DNA* repair associated) is a previously well-established cancer gene involved in homologous recombination, and its germline mutations cause hereditary breast and ovarian cancer.[57] *MMPE* (matrix metalloproteinase E) encodes a secreted calcium-binding phosphoprotein that has been identified as a co-factor of the checkpoint kinase CHK1 and protects cells from DNA damage induced killing.[58] It has also been suggested to serve as a target for sensitizing human tumor cells to radiotherapy or chemotherapy.[59] *DLG3*, also known as *MPP3* (membrane palmitoylated protein 3), encodes a member of the membrane-associated guanylate kinase protein family. Epigenetic inactivation of *MPP3* has been shown to occur frequently during CRC development through promoter hypermethylation.[60] *ZNF66* (zinc finger protein 66) has to our knowledge not been linked to cancer. Whether any of these mutations have contributed to genesis of the respective tumors is unclear, as cancer genomes may contain hundreds to thousands of mutations, and most of the mutations are merely background mutations that do not drive tumorigenesis.[61]

In recent years, MSI has been the target of growing interest due to the associated generation of immunogenic tumor antigens and considerable potential for targeted immunotherapies.[62] In addition to MSI, EMAST has been suggested to cause repeat instability in CRC.[13-16] However, the definition of EMAST has varied between different studies, and to date the studies have relied on the sequencing of a small set of tetranucleotide repeats and mostly in small sample sets.[14,18-20,22,23,63] EMAST CRCs have been suggested to portray a distinct clinicopathological profile, and some studies have found associations between EMAST and a higher histological state, chronic inflammation, and poor prognosis.[15,20,26] Hence, EMAST has been suggested to be a biomarker in CRC.[16,17,27] In our study, which to our knowledge is the first genome-wide study focusing on tetranucleotide

instability, we found no evidence for EMAST as a separate entity. Instead, instability of tetranucleotide repeats was observed in tumors exhibiting MSI, and stochastically in MSS CRCs with higher numbers of any microsatellite mutations. Thus, similar to MSI-L, no evidence was found to support EMAST as a character defining a particular subclass of CRC.

ACKNOWLEDGEMENTS

The authors thank Sini Marttinen, Sirpa Soisalo, Marjo Rajalaakso, Inga-Lill Åberg, Iina Vuoristo, Mairi Kuris, Alison Ollikainen, and Heikki Metsola for technical assistance. This study was supported by grants from the Academy of Finland's Center of Excellence Program 2018-2025, #1312041, as well as grant #1335823, iCAN Digital Precision Cancer Medicine Flagship (320185), The Finnish Cancer Society, The Sigrid Juselius Foundation, Jane and Aatos Erkko Foundation, SYSCOL (an EU FP7 Collaborative Project), Biocentrum Finland, HiLIFE, and Instrumentarium Research Foundation. We acknowledge the computational resources provided by the ELIXIR node, hosted at the CSC-IT Center for Science, Finland.

DATA AVAILABILITY STATEMENT

The raw sequence data used in this study is not available as it contains personally identifiable information, which we do not have consent to distribute. The somatic variants have been deposited in the EGA database under the accession code EGAS00001004710.

ORCID

Johanna Kondelin https://orcid.org/0000-0001-9160-0703
Kristiina Rajamäki https://orcid.org/0000-0001-5151-1220

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
2. Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology*. 2010;138(6):2073-2087.e3.
3. Hemminki A, Peltomäki P, Mecklin JP, et al. Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. *Nat Genet*. 1994;8(4):405-410.
4. Peltomäki P, Aaltonen L, Mecklin JP, de la Chapelle A. A breakthrough in solving the genetic background of colon cancer. *Duodecim*. 1993;109(16):1367-1369.
5. Hamelin R, Chalastanis A, Colas C, et al. Clinical and molecular consequences of microsatellite instability in human cancers. *Bull Cancer*. 2008;95(1):121-132.
6. Boland CR, Thibodeau SN, Hamilton SR, et al. 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(22):5248-5257.
7. Laiho P, Launonen V, Lahermo P, et al. Low-level microsatellite instability in most colorectal carcinomas. *Cancer Res*. 2002;62(4):1166-1170.
8. Network TCGA. The cancer genome atlas network. comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337. https://doi.org/10.1038/nature11252.
9. Kim T-M, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*. 2013;155(4):858-868. https://doi.org/10.1016/j.cell.2013.10.015.
10. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med*. 2016;22(11):1342-1350.
11. Tomlinson I, Halford S, Aaltonen L, Hawkins N, Ward R. Does MSI-low exist? *J Pathol*. 2002;197(1):6-13.
12. Maruvka YE, Mouw KW, Karlic R, et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nat Biotechnol*. 2017;35(10):951-959.
13. Mao L, Schoenberg MP, Scicchitano M, et al. Molecular detection of primary bladder cancer by microsatellite analysis. *Science*. 1996;271(5249):659-662.
14. Haugen AC, Goel A, Yamada K, et al. Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Res*. 2008;68(20):8465-8472.
15. Carethers JM, Koi M, Tseng-Rogenski SS. EMAST is a form of microsatellite instability that is initiated by inflammation and modulates colorectal cancer progression. *Genes*. 2015;6(2):185-205.
16. Carethers JM, Jung BH. Genetics and genetic biomarkers in sporadic colorectal cancer. *Gastroenterology*. 2015;149(5):1177-1190.e3. https://doi.org/10.1053/j.gastro.2015.06.047.
17. Watson MMC, Berg M, Søreide K. Prevalence and implications of elevated microsatellite alterations at selected tetranucleotides in cancer. *Br J Cancer*. 2014;111(5):823-827.
18. Tseng-Rogenski SS, Chung H, Wilk MB, Zhang S, Iwaizumi M, Carethers JM. Oxidative stress induces nuclear-to-cytosol shift of hMSH3, a potential mechanism for EMAST in colorectal cancer cells. *PLoS One*. 2012;7(11):e50616.
19. Tseng-Rogenski SS, Hamaya Y, Choi DY, Carethers JM. Interleukin 6 alters localization of hMSH3, leading to DNA mismatch repair defects in colorectal cancer cells. *Gastroenterology*. 2015;148(3):579-589.
20. Lee S-Y, Chung H, Devaraj B, et al. Microsatellite alterations at selected tetranucleotide repeats are associated with morphologies of colorectal neoplasias. *Gastroenterology*. 2010;139(5):1519-1525.
21. Carethers JM. Microsatellite instability pathway and EMAST in colorectal cancer. *Curr Colorectal Cancer Rep*. 2017;13(1):73-80.
22. Adam R, Spier I, Zhao B, et al. Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of colorectal adenomatous polyposis. *Am J Hum Genet*. 2016;99(2):337-351.
23. Yamada K, Kanazawa S, Koike J, et al. Microsatellite instability at tetranucleotide repeats in sporadic colorectal cancer in Japan. *Oncol Rep*. 2010;23(2):551-561.
24. Hile SE, Shabashev S, Eckert KA. Tumor-specific microsatellite instability: do distinct mechanisms underlie the MSI-L and EMAST phenotypes? *Mutat Res*. 2013;743-744:67-77.
25. Campregher C, Schmid G, Ferk F, et al. MSH3-deficiency initiates EMAST without oncogenic transformation of human colon epithelial cells. *PLoS One*. 2012;7(11):e50541.
26. Devaraj B, Lee A, Cabrera BL, et al. Relationship of EMAST and microsatellite instability among patients with rectal cancer. *J Gastrointest Surg*. 2010;14(10):1521-1528.
27. Torshizi Esfahani A, Seyedna SY, Nazemalhosseini Mojarad E, Majd A, Asadzadeh Aghdaei H. MSI-L/EMAST is a predictive biomarker for metastasis in colorectal cancer patients. *J Cell Physiol*. 2019;234(8):13128-13136.
28. Seshagiri S, Stawiski EW, Durinck S, et al. Recurrent R-spondin fusions in colon cancer. *Nature*. 2012;488(7413):660-664. https://doi.org/10.1038/nature11282.
29. Katainen R, Dave K, Pitkänen E, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet*. 2015;47(7):818-821.

30. Kondelin J, Gylfe AE, Lundgren S, et al. Comprehensive evaluation of protein coding mononucleotide microsatellites in microsatellite-unstable colorectal cancer. *Cancer Res*. 2017;77(15):4078-4088.
31. Cortes-Ciriano I, Lee S, Park W-Y, Kim T-M, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun*. 2017;8:15180.
32. Kondelin J, Salokas K, Saarinen L, et al. Comprehensive evaluation of coding region point mutations in microsatellite-unstable colorectal cancer. *EMBO Mol Med*. 2018;10(9):e8552. https://doi.org/10.15252/emmm.201708552.
33. Palin K, Pitkänen E, Turunen M, et al. Contribution of allelic imbalance to colorectal cancer. *Nat Commun*. 2018;9(1):3664.
34. Cajuso T, Sulo P, Tanskanen T, et al. Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival. *Nat Commun*. 2019;10(1):4022.
35. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*. 2020;578(7793):82-93.
36. Rheinbay E, Nielsen MM, Abascal F, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*. 2020;578(7793):102-111.
37. PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, et al. Genomic basis for RNA alterations in cancer. *Nature*. 2020;578(7793):129-136.
38. Li Y, PCAWG Structural Variation Working Group, Roberts ND, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*. 2020;578(7793):112-121. https://doi.org/10.1038/s41586-019-1913-9.
39. Gerstung M, Jolly C, Leshchiner I, et al. The evolutionary history of 2,658 cancers. *Nature*. 2020;578(7793):122-128.
40. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.
41. Fujimoto A, Fujita M, Hasegawa T, et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res*. 2020;30:24-346. https://doi.org/10.1101/gr.255026.119.
42. Tanskanen T, Gylfe AE, Katainen R, et al. Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients. *Scand J Gastroenterol*. 2013;48(6):672-678.
43. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *New England J Med*. 1998;338(21):1481-1487. https://doi.org/10.1056/nejm199805213382101.
44. Salovaara R, Loukola A, Kristo P, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol*. 2000;18(11):2193-2200. https://doi.org/10.1200/jco.2000.18.11.2193.
45. Weisenberger DJ, Siegmund KD, Campan M, et al. CpG Island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet*. 2006;38(7):787-793.
46. Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol*. 2011;6:479-507.
47. Katainen R, Donner I, Cajuso T, et al. Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer. *Nat Protoc*. 2018;13(11):2580-2600.
48. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer; 2016.
49. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018 https://www.R-project.org/.
50. Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York, NY: Springer; 2002.
51. Watson MM, Lea D, Rewcastle E, Hagland HR, Søreide K. Elevated microsatellite alterations at selected tetranucleotides in early-stage colorectal cancers with and without high-frequency microsatellite instability: same, same but different? *Cancer Med*. 2016;5(7):1580-1587.
52. Haraldsdottir S, Hampel H, Tomsic J, et al. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology*. 2014;147(6):1308-1316.e1.
53. Elsayed FA, Kets CM, Ruano D, et al. Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *Eur J Hum Genet*. 2015;23(8):1080-1084.
54. Jansen AM, van Wezel T, van den Akker BE, et al. Combined mismatch repair and POLE/POLD1 defects explain unresolved suspected lynch syndrome cancers. *Eur J Hum Genet*. 2016;24(7):1089-1092.
55. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.
56. Garcia M, Choi C, Kim H-R, et al. Association between recurrent metastasis from stage II and III primary colorectal tumors and moderate microsatellite instability. *Gastroenterology*. 2012;143(1):48-50.e1.
57. Walsh CS. Two decades beyond BRCA1/2: homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecol Oncol*. 2015;137(2):343-350. https://doi.org/10.1016/j.ygyno.2015.02.017.
58. Liu S, Wang H, Wang X, et al. MEPE/OF45 protects cells from DNA damage induced killing via stabilizing CHK1. *Nucleic Acids Res*. 2009;37(22):7447-7454. https://doi.org/10.1093/nar/gkp768.
59. Sheng J, Luo W, Yu F, Gao N, Hu B. MicroRNA-376a sensitizes cells following DNA damage by downregulating MEPE expression. *Cancer Biother Radiopharm*. 2013;28(7):523-529.
60. Feng X, Chen K, Ye S, et al. MPP3 inactivation by promoter CpG islands hypermethylation in colorectal carcinogenesis. *Cancer Biomark*. 2012;11(2–3):99-106.
61. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546-1558.
62. Samstein RM, Lee C-H, Shoushtari AN, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet*. 2019;51(2):202-206.
63. Hamaya Y, Guarinos C, Tseng-Rogenski SS, et al. Efficacy of adjuvant 5-fluorouracil therapy for patients with EMAST-positive stage II/III colorectal cancer. *PLoS One*. 2015;10(5):e0127591.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kondelin J, Martin S, Katainen R, et al. No evidence of EMAST in whole genome sequencing data from 248 colorectal cancers. *Genes Chromosomes Cancer*. 2021;60:463–473. https://doi.org/10.1002/gcc.22941