Analysis of Somatic Microsatellite Indels Identifies Driver Events in Human Tumors

Ed sum:
New computational tools reveal the contribution of microsatellite insertions and deletions to the mutational landscape in human cancer.

Analysis of somatic microsatellite indels identifies driver events in human tumors

Yosef E. Maruvka[1,2,3], Kent W. Mouw[4,5], Rosa Karlic[6], Prasanna Parasuraman[1,2], Atanas Kamburov[1,2,3], Paz Polak[1,2,3], Nicholas J. Haradhvala[1,2,3], Julian M. Hess[3], Esther Rheinbay[1,2,3], Yehuda Brody[3], Amnon Koren[7], Lior Z Braunstein[1,2,3], Alan D'Andrea[3,4,5], Michael S Lawrence[1,2,3], Adam Bass[3,8], Andre Bernards[1,2], Franziska Michor[9], Gad Getz[1,2,3,4]

[1] Massachusetts General Hospital Center for Cancer Research, Charlestown, Massachusetts 02129, USA

[2] Massachusetts General Hospital, Department of Pathology, Boston, Massachusetts 02114, USA

[3] Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA

[4] Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA

[5] Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[6] Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia

[7] Cornell University Department of Molecular Biology and Genetics, 526 Campus Road, Ithaca, NY 14853, USA

[8] Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[9] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.

Corresponding Author:
Gad Getz, PhD
Cancer Program, Broad Institute of MIT and Harvard
415 Main St.
Cambridge, MA 02142
gadgetz@broadinstitute.org

Abstract

Microsatellites (MSs) are tracts of variable-length repeats of short DNA motifs that exhibit high rates of mutation in the form of insertions or deletions (indels) of the repeated motif. Despite their prevalence, the contribution of somatic MS indels to cancer is largely unexplored due to difficulties in detecting them in short-read sequencing data. Here we present two tools: MSMuTect, for accurate detection of somatic MS indels, and MSMutSig, for identification of genes containing MS indels at higher frequency than expected by chance. Applying MSMuTect to whole-exome data from 6,747 human tumors representing 20 tumor types, we identified >1000 novel MS indels in cancer genes. Additionally, we demonstrate that the number and pattern of MS indels can accurately distinguish microsatellite stable (MSS) tumors from tumors with microsatellite instability (MSI), which may improve classification of clinically relevant subgroups. Finally, we identify seven novel MS indel driver hotspots – four in known cancer genes (*ACVR2A*, *RNF43, JAK1,* and *MSH3*) and three in genes not previously implicated as cancer drivers (*ESRP1*, *PRDM2* and *DOCK3*).

Introduction

Microsatellites (MSs), are regions of the genome characterized by repetition of a short sequence motif (usually 1-6 bp)[1]. MSs are abundant in non-transcribed regions of the human genome, but also occur in exons and untranslated regions (UTRs; **Fig. S1**). In the germline, rates of insertions and deletions (indels) in MSs are significantly higher than rates of single nucleotide substitutions elsewhere in the genome (10^{-4}-10^{-3} compared to ~10^{-8} per locus per generation, respectively)[2]. The increased mutation rate within MS indels is thought to arise due to DNA polymerase slippage during replication, leading to changes in the number of repeats. MS indels frequently result in frameshift mutations and can therefore dramatically alter protein function or expression[1].

More than 40 hereditary diseases are caused by germline MS indels[3,4,5]. In addition, many cancer genes[6] (e.g. *PTEN* and *NF1*) contain MS loci, and in some cases, somatic MS indels have been causally implicated in cancer[7]. Tumors with microsatellite instability (MSI) have dramatically increased numbers of MS indels owing to loss of normal mismatch repair (MMR) function[8]. Although the MSI phenotype has been observed across tumor types, it appears to be most common in colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC)[8]. Given the important prognostic and therapeutic implications of MSI status, many clinical centers perform PCR- or immunohistochemistry-based MSI testing for these tumor types [9–12].

Despite their potential significance, somatic MS indels have not been systematically analyzed in cancer due to challenges associated with their detection via current massively parallel sequencing data, including read length limits and PCR errors.[13] The frequency of such sequencing errors varies significantly across MS loci (**Online Methods**); therefore, methods utilizing principled statistical modeling and noise estimation are required to accurately identify MS indel events.

Discovering cancer-associated MS loci relies on identifying evidence of positive selection (i.e., mutation frequencies higher than expected by chance). However, simply comparing the frequency of mutations at each MS locus to the average mutation frequency across the genome is inadequate, as the background mutation frequency can vary by nearly two orders of magnitude[14]. Therefore, accurate estimates of site-specific background mutation frequencies are required to maximize the sensitivity to discover cancer-associated MS loci while minimizing the rate of false calls[15].

To address these challenges, we developed two new tools: MSMuTect for detecting somatic MS indels from sequencing data and MSMutSig for detecting loci and genes with significantly increased frequency of MS indels. Applying these tools across 6,747 tumors representing 20 tumor types, we uncover unique properties of MS indels and identify MS loci that likely represent cancer driver events. Comparison of MS indels across clinical MS groups reveals differences between MSS and MSI tumors that may be relevant for clinical decision-making.

Results

MSMuTect identifies MS indels from exome sequencing data

In an effort to improve detection of somatic MS indels, we globally re-aligned reads[7] from whole-exome sequencing data of 6,747 tumor/normal pairs across 20 tumor types from The Cancer Genome Atlas (TCGA) to the unique sequence flanking 383,515 MS loci, defined as sites with at least five repeats of a 1-6 bp motif in the exome territory (**Online Methods**; **Fig. S2**)[1]. This re-alignment step reduced the fraction of misaligned reads compared to the standard alignment (**Fig. S3**). We then counted, for every MS locus, the number of reads supporting each MS repeat length, thus producing two histograms of MS repeat lengths, one for the tumor and one for the matched normal sample (**Fig. 1A**).

Sequencing errors, PCR amplification errors, and other sources of noise can change the number of MS repeats present in a given read. Therefore, the true underlying allele(s) must be statistically inferred from the data (**Fig. 1B**). Critical to this inference is the empirical estimation of the noise associated with each type of MS. We trained empirical noise models, one for each MS type (defined by its motif and number of repeats), using data from homozygous sites derived from the X chromosome of 4,411 male normal samples (i.e., sites having only one true allele at each MS locus). We had sufficient data to reliably estimate the noise models for the motifs A, C, AC, and AG, which together represent 98% of the MS loci in the exome (**Online Methods**).

To accurately identify the alleles present in tumor and normal samples and detect somatic MS indels, we used these noise models to calculate, for each MS locus, the set of most likely alleles (**Fig. 1C**; **Online Methods**). For each locus, we used a log-likelihood ratio test to compare models in which the locus harbored one versus two distinct alleles (either distinct germline alleles or a somatic mutation at a homozygous site). If the two-allele model fit the data better, we then compared it to a three-allele model, and so forth, to a maximum of four alleles. Finally, in cases where the set of alleles were different between the tumor and normal, we reported a somatic MS indel event only after ensuring that the histogram of MS repeat lengths in the tumor was indeed described better by the inferred tumor alleles than by the inferred normal alleles (**Fig. 1C; Online Methods**).

We tested the sensitivity and specificity of MSMuTect using an approach similar to that previously described for MuTect[17]. We analyzed sequencing replicates from a single individual and selected parameters that, on average, generated no greater than five false positives per exome (**Fig S4; Online Methods**). To evaluate sensitivity, we simulated MS indels by inserting or deleting a single motif repeat at different loci throughout the genome. We evaluated the sensitivity across various allele fractions and MS loci lengths and found that it was highest for shorter MS loci and decreased when the MS indel allele fraction fell below 20% (**Fig. S5; Online Methods**).

MS indel mutational landscape

We applied MSMuTect across 6,747 TCGA whole exome tumor/normal pairs representing 20 tumor types (**Tables S1**, **S2**). Our analysis identified 174,638 MS indels, with a range of 0 to 900 per tumor. We observed extensive inter- and intra-tumor variability in the MS indel rate, similar to the variability reported for single nucleotide variations and copy-number alterations (**Fig. 2; Fig S6**)[15]. The average MS indel frequency varied significantly across tumor types, with the highest frequencies in colorectal (COAD, READ), stomach (STAD) and endometrial tumors (UCEC), consistent with frequent MMR deficiency in these tumors.

Breast cancer (BRCA) had the fifth highest MS indel rate, and although BRCA is not typically thought to have high rates of MS indels, there is a subset with known MSI features[19], and a recent study[6] identified mutational signatures consistent with loss of mismatch repair in BRCA. A recent report by Hause et al[13] did not identify MSI-H cases in TCGA BRCA; however, they analyzed a subset of the TCGA breast cancer cohort (266/1069 cases) that included only 3/36 tumors in which we identified >100 MS indels. Previous reports[20,21,22] have identified a small fraction of MSI cases in cohorts of cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS) and adrenocortical carcinoma (ACC), and indeed our analysis identified MS indels in these tumor types (ranked 6th, 7th and 8th in average MS indel frequency, respectively).

To validate the identified MS indels, we analyzed RNA-seq data available for a subset of the samples (**Table S3**). For each of the 150 significantly mutated MS indels (described below) with sufficient RNA-seq coverage (≥4 reads), we manually compared the alleles inferred by MSMuTect to the alleles observed in the RNA-seq data, and validated 87% of them (**Table S3**). Importantly, RNA-seq likely underestimates the accuracy of MSMuTect because MS indels that introduce premature stop codons can trigger nonsense-mediated decay of the altered mRNA transcript, thus decreasing the likelihood of observing RNA-seq reads that support the MS indel. Indeed, MS loci closer to the 3' end of the transcript, which are less likely to trigger nonsense-mediated decay, had higher validation rates (e.g. ACVR2A (96%) and RNF43 (100%); **Table S3)**. For four of the five cases in which two distinct somatic events were identified at the same site, we were able to validate all three alleles (one wild type and two alternate alleles).

MSMuTect correctly classifies tumors with respect to MS stability

We next asked whether MSMuTect could recapitulate independent measures of tumor MS stability. As part of TCGA, tumors from the COAD, STAD, and UCEC cohorts were experimentally classified as exhibiting microsatellite stability (MSS, no indels) or microsatellite instability (MSI-low [MSI-L], indel at one MS locus; MSI-high [MSI-H], indels at 2 or more MS loci) using a PCR-based assay to assess size variability at the five Bethesda MS loci[12]. Applying MSMuTect, we found that tumors classified as MSI-H had significantly more MS indels than samples that were MSS or MSI-L (MSI-H vs MSS: COAD median 104.5 vs. 3.0, $P < 10^{-22}$; STAD 64.5 vs. 1.0, $P < 10^{-28}$; UCEC 94.5 vs. 2.0, $P < 10^{-58}$, Mann-Whitney; **Fig. 3A; Online Methods**). There was no difference in the number of MS indels in tumors classified as MSI-L versus MSS for COAD, but there was a small difference in UCEC and STAD (UCEC median 9 vs. 2, $P < 10^{-6}$; STAD 3 vs 2, $P < 10^{-3}$, Mann-Whitney) due to contribution from a small number of MSI-L cases with many MS indels (discussed below). In addition, we found that MSI-H tumors were significantly ($P < 10^{-16}$, t-test, **Fig. S7**) more likely to have several MS indels at the same locus and were also more likely to have one (or more) MS indels at heterozygous germline MS sites.

Although MSMuTect separates the majority of MSI-H tumors from MSI-L and MSS tumors, there were several cases with an apparent discrepancy between the MS indel count and the Bethesda designation (**Fig. 3A**). MMR-deficient tumors are known to have a specific pattern of SNVs (MSI-SNV signature), and thus the fraction of SNVs associated with the MSI-SNV signature can be used as an orthogonal metric to identify the MSI phenotype[23](**Online Methods**). As expected, tumors in which MSI-SNVs comprised >15% of the total SNVs (red in **Fig. 3A**) were nearly all (264/277) classified as MSI-H and had high MS indel counts. In addition, 7 of the 12 MSI-H STAD and UCEC tumors with the lowest MS indel counts (<10) had an MSI-SNV fraction <15% (blue in **Fig. 3A**), suggesting that the samples may have been misclassified as MSI-H by the PCR-based assay.

We also observed that many of the MSI-L and MSS samples with the highest number of MS indels also had a relatively high number of total SNVs (**Fig. 3A**). Mutations in the exonuclease (proofreading) domain of polymerase epsilon (POLE) can dramatically increase the number of SNVs; therefore, to investigate the potential interaction of POLE-mediated mutagenesis with MS indels, we calculated the fraction of SNVs that were likely contributed by POLE-mediated mutagenesis (POLE-SNVs; **Online Methods**). All but one of the 63 samples in which POLE-SNVs comprised >15% of the total SNVs had a somatic missense mutation in the exonuclease domain of POLE (n=60) or polymerase delta (*POLD1*; n=2). Although the majority of the POLE/POLD1-mutated tumors (45/63) were classified as MSS or MSI-L, they had significantly more MS indels than other MSS and MSI-L tumors (median 54 vs. 2 among MSI-L and 18.5 vs 2 among MSS), raising the possibility that POLE/POLD1 exonuclease domain mutations may contribute to the MS indel burden and highlighting the limitations of the PCR-based MSI assay (**Fig. 3A**).

Differences in MS indel properties in MSS and MSI samples

In addition to differences in numbers of MS indels, MSI and MSS samples also differ in their association between MS indel frequency and DNA replication timing, and both are distinct from the association reported for SNVs[14,24]. In general, unlike SNV density, MS indel density does not show a strong correlation with replication timing. Interestingly, there was no correlation in MSS samples (slope = -0.03, Pearson correlation = -0.47, *P*=0.43, t-test; **Fig**. **3B**), and only a marginal, but significant, decrease with replication timing in MSI samples (slope = -0.1, Pearson correlation = -0.995, *P*=0.0003, t-test; **Fig. 3B**)[25], which is opposite to the direction observed for SNVs.

Likewise, in both MSI and MSS tumors, MS indels are more common at loci with longer repeat lengths; however, the slope and shape of these relationships differ **(Fig. 3C)**. Moreover, the ratio of insertions to deletions is different between MSS and MSI cases, with MSI cases having a tendency towards deletions[25] while MSS cases tend towards insertions (**Fig. 3D**; $P<10^{-31}$, χ^2 test). The tendency to increase repeat lengths in MSS cases is consistent with germline MS indels, which have been shown to preferentially undergo insertions in MS loci with <15 repeats[2].

MS indels in known cancer genes

We next sought to identify somatic MS indels that drive tumorigenesis. We first attempted to identify novel MS indels across 727 known cancer genes[6] in a cohort of 4,064 TCGA samples with curated mutations calls (**Online Methods**). We focused our analysis on MS loci for which the inferred allele matched the reference allele in at least 90% of the normal (germline) samples (**Fig. S8**). MS indels at loci with greater germline diversity may have weaker functional effects or represent noisy sites. We detected 1470 MS indels across these genes (**Table S4**), including 89 indels that had been previously identified by the TCGA consortium and 1105 indels in samples without any other indel or non-synonymous SNVs reported in the same gene (thus potentially representing novel loss-of-function events in these genes). The remaining 276 indels were identified in samples that had a separate event (indel or nonsynonymous SNV) in the same gene; in these cases, the identified MS indel may represent the "second hit"[26]. In some genes, previously unidentified MS indels comprise a substantial fraction of the total number of mutations. Reassuringly, MS indels were enriched in tumor suppressor genes (TSGs)[27] compared to oncogenes (993 MS indels in 70 TSGs vs 272 MS indels in 53 oncogenes, $P<10^{-58}$, binomial test).

MSMutSig, A Tool for Identifying Driver MS indels

Next, we extended our MutSig suite of tools[15] for detecting candidate cancer genes and developed MSMutSig to specifically address the unique properties of MS indels. Our analysis of ~250,000 MS loci revealed that the two major factors (covariates) that influence the mutation frequency of an MS locus are the motif sequence and repeat length (**Fig. 3C; Online Methods**). We, therefore, estimated the background mutation frequency for each motif and repeat length separately. We first attempted to apply a binomial model but found that many loci harbored more (or fewer) mutations than predicted by the model (**Fig. S9**). To address this, we applied a more dispersed distribution – the negative-binomial – and fit the extra dispersion parameter such that no MS loci in non-coding regions would be nominated as significantly mutated (at Benjamini-Hochberg FDR q<0.1). This model indeed also captured the variability of MS indel rates in coding regions (**Figs. S10-S12**).

Once optimized, we applied MSMutSig across 20 tumor types. For the three tumor types with high frequencies of MSI cases (COAD, STAD, and UCEC), we considered the MSS and MSI subgroups (as defined by TCGA) separately (**Fig. 3B-D**). The only tumor types that yielded significant MS loci (q<0.1) were the MSI subtypes of COAD, STAD, and UCEC. In COAD, we identified 3 significant MS loci in the genes *ACVR2A*, *RNF43* and *DOCK3*; in STAD, 4 loci in *ACVR2A*, *RNF43*, *MSH3* and *PRDM2*; and in UCEC, 5 loci in *RNF43*, *DOCK3*, *JAK1*, *ESRP1* and *ACVR2A*. Thus, our analysis nominated a total of 7 MS hotspots in 7 genes (**Table 1**). Three of these genes (*ACVR2A*, *RNF43* and *JAK1*) have been previously identified as cancer genes based on an increased mutation frequency in one or more tumor types[6] (**Fig. 5**). In the TCGA colon cancer study[28], *MSH3* was not nominated as significantly mutated but was noted to be highly mutated by manual examination of the sequence data. Notably, due to the high mutability of MS loci, beyond major cancer genome studies, the literature is mixed regarding which of the 17,398 genes with MS loci are associated with cancer.[29] The remaining three genes (*ESRP1*, *PRDM2* and *DOCK3*) have not been previously identified as cancer genes (discussed below). Previously identified cancer drivers such as *TGFRB2*[30] and *RPL22*[31] are missing from our list since their MS loci were excluded from the analysis due to high variability in germline samples.

All seven of the significantly mutated MS indels cause a frameshift mutation within an exon. Frameshift mutations typically result in decreased gene expression because the altered mRNA undergoes nonsense-mediated decay[32]. However, if a frameshift mutation occurs near the end of a gene, nonsense-mediated decay is less likely to occur[33]. Of the 7 MS indels identified here, four (in *ESRP1*, *MSH3*, *JAK1*, and *PRDM2*) lead to a significant reduction in mRNA expression levels (**Table 1**, **Fig. 4**) whereas the MS indels in *ACVR2A* and *DOCK3* occur near the 3' end of the gene and thus are not expected to lead to nonsense-mediated decay. The MS indel in *RNF43* is in the second to last exon; however, the presence of this indel does not correlate with a reduced *RNF43* expression level (Mann-Whitney *P*=0.4), and may represent an exception to the '50 bp rule', similar to *UPF1*.[34,35]

Genes Nominated by MSMutSig are Candidate Cancer Drivers

The *ACVR2A* gene, encoding Activin A Receptor Type IIA, harbors the most frequently mutated novel MS locus in our list (p.K437fs), with mutations in ~80% (32/40) of MSI colon tumors, ~75% (52/69) of MSI stomach tumors, and ~19% (29/157) of MSI endometrial tumors (**Table 1; Fig 5**). *ACVR2A* is a member of the TGF-β signaling pathway, which plays a major role in cell growth and is known to be highly mutated in all three of these tumor types. Consistent with a tumor suppressor role, two studies [36,37] showed that expression of wildtype *ACVR2A* in MSI colon cancer cell lines with mutated *ACVR2A* led to reduced cell growth. When these novel MS indel events are considered with other reported alterations, *ACVR2A* is among the most frequently mutated genes in colorectal cancer with mutations in ~20% of all cases.

The gene encoding Ring Finger Protein 43 (*RNF43*) harbors the MS indel p.G659fs in 40% (16/40) of MSI colon tumors, 35% (24/69) of MSI stomach tumors, and 23% (36/155) of MSI endometrial tumors. *RNF43* is a negative regulator of the WNT signaling pathway, which is involved in controlling cell proliferation[38]. This gene was reported as significant in STAD by TCGA[39], however it was not due to its MS indels. Giannakis et al.[7] recently detected the same *RNF43* MS indel through manual review of *RNF43* sequence data and determined that it is frequently present in colon and endometrial tumors.

The gene encoding the protein MutS Homolog 3 (*MSH3*) is a member of the Mismatch Repair (MMR) pathway, and germline mutations in *MSH3* are known to increase the risk of developing MSI tumors[41]. We identified the MS indel hotspot p.K383fs in 40% (28/69) of stomach tumors. In mouse models, inactivation of *MSH3* alone does not lead to cancer, but concurrent loss of *MSH3* and *MSH6* results in an increased rate of tumor formation[42].

The gene encoding the protein PR domain 2 (*PRDM2*), is a histone H3 lysine 9 methyltransferase and has been implicated as a tumor suppressor in several tumor types[43]. Decreased *PRDM2* expression has been associated with renal cell carcinoma[44], esophageal squamous cell carcinoma[45] and meningiomas[46]. We identified the MS indel hotspot p.K1489fs in 48% (33/69) of stomach tumors. Analysis of gene expression data revealed decreased expression in mutated cases (*P*=0.016 Mann-Whitney; **Fig. S13**), consistent with partial nonsense-mediated decay.

The epithelial splicing regulatory protein 1 (*ESRP1*) is an epithelial cell-type-specific splicing regulator[47]. Its MS indel hotspot (*ESRP1* p.K511fs) is mutated in approximately 20% (31/158) of MSI endometrial tumors. *ESRP1* regulates alternative splicing of *FGFR2*[47] from the IIIc mesenchymal isoform to the IIIb epithelial isoform[47]. Thus, mutations in *ESRP1* may contribute to the epithelial-mesenchymal transition (EMT). In pancreatic cancer[48], the transition from expression of the FGFR2-IIIb isoform to the FGFR2-IIIc isoform is associated with increased cell growth, migration, and invasion. We analyzed TCGA RNA-seq data and found that MS indels in *ESRP1* were associated with both a significant decrease in *ESRP1* expression (**Fig. 4a**; $P < 1.5\text{x}10^{-9}$ Mann-Whitney) as well as a significant increase in the ratio of isoform IIIc to IIIb in *ESRP1* mutant cases (**Fig. 4b**; $P < 9\text{x}10^{-7}$ Mann-Whitney).

Our finding that *JAK1* harbors the frameshift mutation p.N860fs in 21% (33/158) of endometrial tumors (**Table 1**) was somewhat unexpected given *JAK1*'s known role as an oncogene[49]. Park et al.[25] found that the *JAK1* p.N860fs indel is associated with repression of transcript levels of *JAK1* downstream targets, and a recent study[50] suggested that truncated *JAK1* modulates the IFNγ signaling pathway and enables tumor immune evasion. We compared expression of an IFNγ-mediated gene signature[51] in tumors with or without the *JAK1* p.N860fs indel and found a significant reduction in expression in 21 of 27 IFNγ-related genes in tumors with the p.N860fs indel. Therefore, *JAK1* loss may promote tumor survival by inhibiting an IFNγ-mediated antitumor immune response.

Finally, *DOCK3* encodes the protein Dedicator of cytokinesis 3 and carries the MS indel mutation p.T1850fs in 40% of colon tumors (16/40) and 23% of endometrial tumors (33/145) (**Table 1**). *DOCK3* (also known as *MOCA*) is an exchange factor for Rac GTPases and was recently implicated as an inhibitor of the WNT signaling pathway[52]. *CTNNB1*, a core member of the WNT pathway, is mutated in approximately 30% of endometrial tumors, and *DOCK3* mutations are mutually exclusive with *CTNNB1* mutations ($P < 0.015$, hypergeometric test in

UCEC MSI cases; *P*<0.005 among all UCEC cases).

Discussion

Here, we introduce MSMuTect, a tool for accurately identifying somatic indels in MS loci, and MSMutSig, a tool for identifying candidate cancer genes with significantly enriched MS indel events. MSMuTect relies on careful realignment of MS-containing reads to MS loci and uses a principled statistical test to identify somatic events by applying an empirical noise profile based on motif and repeat length. Given the wide variation in background mutation rates across MS loci, this approach is necessary to reduce the rate of false positive MS indel calls.

An alternate method for detecting somatic MS indels, recently reported by Hause et al[13], nominates a somatic MS indel if a single tumor read supports a different number of motif repeats than the normal sample. This approach for calling MS indels results in a large number of apparent MS indels, with a median of ~900 MS indels per MSS sample compared to <10 using MSMuTect, and only a ~3-fold difference in number of MS indels between MSS and MSI cases (897 in MSS vs 3009 in MSI) versus an ~18-fold difference using MSMuTect (8 vs 145). These apparent differences may be due in part to the inclusion of many subclonal MS indels by Hause et al, whereas MSMuTect primarily considers clonal events.

MSMuTect infers the alleles in both the tumor and normal (germline) samples, and somatic MS indels are nominated only when the observed MS repeat lengths in the tumor are better explained by the tumor allele(s) than by the normal allele(s). A recent report by Kim et al[25] used the Kolmogov-Smirnov (KS) test to compare repeat length distributions in the tumor and normal sample at each MS locus for a limited set of endometrial and colon cancers. Although the total number of reported MS loci is comparable to the number identified by MSMuTect (median of

~150 for MSI-H cases and ~2 for MSS), the KS test does not infer the actual (potentially multiple) alleles in the tumor and normal samples. In addition to reducing the risk of false-positive MS indel calls, identifying MS alleles in the normal sample has the potential to discover novel germline MS indels. Indeed, we found that a small percentage of cases (5/6748, 0.075%) had a germline *RNF43* allele that was identical to the most common somatic *RNF43* mutant allele, raising the possibility of an inherited pathogenic *RNF43* MS indel. A recent study by Taupin et al.[53] showed that inherited *RNF43* variants are a risk factor for the familial cancer syndrome Serrated Polyposis.

We applied MSMuTect across 6,747 cases representing 20 tumor types from the TCGA dataset and identified nearly 175,000 MS indels. As expected, the tumor types with the highest rates of MS indels were those classically associated with the MSI phenotype – colon, rectal, stomach, and endometrial tumors. However, several other tumor types – including breast and cervical cancers – had a notable percentage of cases with high numbers of MS indels, suggesting that the MSI phenotype also occurs in these tumor types and that MSI testing may be warranted for these tumors in certain clinical settings, such as when screening for immunotherapy trials[10,54,55].

Comparing traditional PCR-based stratification of the MSI status of the TCGA COAD, STAD, and UCEC cohorts with our results, we found a significant difference in MS indel frequency between MSI-H and both MSI-L and MSS tumors, but the difference between MSI-L and MSS tumors was less significant. Furthermore, there was significant variability in MS indel frequency within each MS subgroup, particularly among endometrial tumors.

We found that many MSI-L tumors have MS indel frequencies similar to those of MSS tumors, suggesting that some were misclassified by the MSI assay. However, a subset of MSI-L tumors – particularly endometrial tumors – have MS indel frequencies that more closely resemble MSI-H tumors. Proofreading deficient tumors arising from POLE/POLD1 exonuclease mutations have dramatically increased rates of SNVs and a characteristic mutational signature, as well as high MS indel rates, including in MSI-L and MSS cases (**Fig. 3a**). Thus, whereas some endometrial tumors appear to have concomitant POLE mutations and MSI, other MSI-L endometrial tumors have an SNV signature consistent with a POLE/MSS phenotype despite their relatively high number of MS indels. To our knowledge, an interaction between somatic POLE mutations and MS indels has not been reported, although a similar association was recently noted in yeast[56]. It is possible that a dramatic increase in SNVs resulting from loss of POLE proofreading may saturate MMR capacity (which corrects both SNVs and indels) and thus indirectly result in an increased number of unrepaired MS indels.

Finally, we noted that many of the MSI-H/MSI-L endometrial tumors with lowest MS indel frequencies lacked the MSI-SNV signature, suggesting that these tumors may have been misclassified as MSI, further underscoring the differences between the PCR-based MSI assay and MSI classification derived from whole exome sequencing. These results highlight the limitations of clinical MSI assays and, given the recent identification of MSI as a biomarker of immunotherapy response, underscore the need for sensitive and reliable MSI assays[10,54].

Based on our understanding of the features that influence the indel mutation rate at MS loci, we developed MSMutSig, a tool that identifies MS loci that are mutated more frequently than expected by chance. MutSig[15] was developed to handle SNVs and general indels (not necessarily within MSs), and its background mutation rate model does not fit the unique properties of MS indels. Indeed, applying MutSig to the MSI-H endometrial cohort using all mutations (SNVs, general indels and MS indels) yielded 296 significant genes ($q<0.1$) and an inflated Q-Q plot, suggesting an inadequate null model (**Fig. S14**).

Many genes have been proposed to drive cancer based on a high frequency of MS indels[29]; however, our analysis suggests that many of these MS loci have a high background mutation rate and therefore may be frequently mutated but not under selective pressure. Applying MSMutSig across 6,747 cases, we identified seven significantly mutated MS loci, three of which occurred in genes not previously nominated as cancer genes (*ESRP1*, *PRDM2* and *DOCK3*). Although our analysis strongly supports a role for these MS indels as cancer drivers, direct experimental studies will be needed to further investigate the specific role of these mutations in cancer.

In these analyses, we assumed that all MS indels in non-coding regions were not under selective pressure and thus could serve as an upper estimate of the indel mutation rate arising from technical factors (PCR errors, etc). However, cancer driver mutations are known to exist in regulatory regions[57,58,59], and a deeper understanding of the covariates influencing MS indel rates across the genome may eventually enable us to adapt MSMutSig for accurate detection of significantly mutated MS loci in non-coding regions. Similarly, adapting MSMuTect for whole genome analysis may further improve the sensitivity of MSMuTect and MSMutSig by providing a more accurate noise model across loci of varying motif and repeat lengths. In addition, technical advances may also lead to improvements in MS indel calling. For example, sequencing technologies that produce longer read lengths will provide better coverage of long MS loci and enable more accurate mutation-calling for longer MS repeats. Finally, integrating MS indel calling tools such as MSMuTect with tools for identifying other recurrent genomic events such as SNVs or copy number alterations (CNAs) will provide a more comprehensive view of cancer driver events.

Acknowledgments

We thank C. Mayer for supplying and supporting the phobos tool. G. Getz was partially funded by the Paul C. Zamecnick, MD, Chair in Oncology at MGH and the NIH TCGA Genome Data Analysis Center (NIH U24CA143845). Y. Maruvka, P. Polak, and A. Kamburov were funded by the G. Getz start-up funds at Massachusetts General Hospital. K. Mouw was partially funded by an American Society of Radiation Oncology (ASTRO) Junior Faculty Career Research Training Award and a Harvard Catalyst KL2/CMeRIT Award. F. Michor gratefully acknowledges support from the Dana-Farber Cancer Institute Physical Sciencies Oncology Center (NIH U54CA193461).

Author Contributions

Y.E.M., K.W.M., F.M., and G.G. devised the research strategy. Y.E.M. and G.G. developed the tools. Y.E.M., R.K., N.J.H., and J.M.H. performed analyses. Y.E.M, K.W.M, R.C., Pr.P., A.K., Pa,P,N.J.H, J.M.H., E.R., Y.B., A.K., L.Z.B, A.D., M.S.L, Ad.B., An.B., F.M., and G.G. helped interpret results. Y.E.M., K.W.M., and G.G. wrote the manuscript. All authors reviewed and approved the final manuscript.

The authors declare no competing financial interests.

References

1. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5,** 435–445 (2004).
2. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44,** 1161–1165 (2012).
3. Pearson, C. E., Edamura, K. N. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6,** 729–742 (2005).
4. Kennedy, L. *et al.* Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum. Mol. Genet.* **12,** 3359–3367 (2003).
5. Willemsen, R., Levenga, J. & Oostra, B. A. CGG repeat in the FMR1 gene: size matters. *Clin. Genet.* **80,** 214–225 (2011).
6. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (2016).
7. Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46,** 1264–1266 (2014).
8. Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer—the stable evidence. *Nat. Rev. Clin. Oncol.* **7,** 153–162 (2010).
9. Stadler, Z. K. Diagnosis and management of DNA mismatch repair-deficient colorectal cancer. *Hematol. Oncol. Clin. North Am.* **29,** 29–41 (2015).
10. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372,** 2509–2520 (2015).
11. Watkins, J. C. *et al.* Universal Screening for Mismatch-Repair Deficiency in Endometrial Cancers to Identify Patients With Lynch Syndrome and Lynch-like Syndrome. *Int. J. Gynecol. Pathol. Off. J. Int. Soc. Gynecol. Pathol.* (2016). doi:10.1097/PGP.0000000000000312

12. Umar, A. *et al.* Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability. *J. Natl. Cancer Inst.* **96,** 261–268 (2004).

13. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22,** 1342–1350 (2016).

14. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

15. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

16. Mayer, C., Leese, F. & Tollrian, R. Genome-wide analysis of tandem repeats in Daphnia pulex-a comparative approach. *BMC Genomics* **11,** 277 (2010).

17. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

18. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

19. Tokunaga, E. *et al.* Frequency of microsatellite instability inBreast cancer determined by high-resolution fluorescent microsatellite analysis. *Oncology* **59,** 44–49 (2000).

20. Larson, A. A. *et al.* Analysis of replication error (RER+) phenotypes in cervical carcinoma. *Cancer Res.* **56,** 1426–1431 (1996).

21. Taylor, N. P. *et al.* Defective DNA mismatch repair and XRCC2 mutation in uterine carcinosarcomas. *Gynecol. Oncol.* **100,** 107–110 (2006).

22. Medina-Arana, V. *et al.* Adrenocortical carcinoma, an unusual extracolonic tumor associated with Lynch II syndrome. *Fam. Cancer* **10,** 265–271 (2011).

23. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).

24. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4,** 1502 (2013).

25. Kim, T.-M., Laird, P. W. & Park, P. J. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155,** 858–868 (2013).

26. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci.* **68,** 820–823 (1971).

27. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558 (2013).

28. Network, T. C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).

29. Cederquist, K. Genetic and epidemiological studies of hereditary colorectal cancer. (2005).

30. Biswas, S. *et al.* Mutational inactivation of TGFBR2 in microsatellite unstable colon cancer arises from the cooperation of genomic instability and the clonal outgrowth of transforming growth factor β resistant cells. *Genes. Chromosomes Cancer* **47,** 95–106 (2008).

31. Network, T. C. G. A. R. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).

32. Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5,** 89–99 (2004).

33. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci.* **100,** 189–192 (2003).

34. Zhang, J., Sun, X., Qian, Y. & Maquat, L. E. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA N. Y. N* **4,** 801–815 (1998).

35. Silva, A. L. *et al.* The canonical UPF1-dependent nonsense-mediated mRNA decay is inhibited in transcripts carrying a short open reading frame independent of sequence context. *RNA* **12,** 2160–2170 (2006).

36. Deacu, E. *et al.* Activin Type II Receptor Restoration in ACVR2-Deficient Colon Cancer Cells Induces Transforming Growth Factor-β Response Pathway Genes. *Cancer Res.* **64,** 7690–7696 (2004).

37. Ballikaya, S. Activin Receptor Type 2 A (ACVR2A)-dependent Proteomic and Glycomic Alterations in a Microsatellite Unstable (MSI) Colorectal Cancer Cell Line Model System. (2014).

38. Niu, L. *et al.* RNF43 Inhibits Cancer Cell Proliferation and Could be a Potential Prognostic Factor for Human Gastric Carcinoma. *Cell. Physiol. Biochem. Int. J. Exp. Cell. Physiol. Biochem. Pharmacol.* **36,** 1835–1846 (2015).

39. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513,** 202–209 (2014).

40. Jo, Y. S. *et al.* Frequent frameshift mutations in 2 mononucleotide repeats of RNF43 gene and its regional heterogeneity in gastric and colorectal cancers. *Hum. Pathol.* **46,** 1640–1646 (2015).

41. Duraturo, F. *et al.* Association of low-risk MSH3 and MSH2 variant alleles with Lynch syndrome: probability of synergistic effects. *Int. J. Cancer* **129,** 1643–1650 (2011).

42. Wind, N. de *et al.* HNPCC-like cancer predisposition in mice through simultaneous loss of Msh3 and Msh6 mismatch-repair protein functions. *Nat. Genet.* **23,** 359–362 (1999).

43. Mzoughi, S., Tan, Y. X., Low, D. & Guccione, E. The role of PRDMs in cancer: one family, two sides. *Curr. Opin. Genet. Dev.* **36,** 83–91 (2016).

44. Ge, P., Yu, X., Wang, Z.-C. & Lin, J. Aberrant Methylation of the 1p36 Tumor Suppressor Gene RIZ1 in Renal Cell Carcinoma. *Asian Pac. J. Cancer Prev.* **16,** 4071–4075 (2015).

45. Dong, S.-W. *et al.* Alteration in gene expression profile and oncogenicity of esophageal squamous cell carcinoma by RIZ1 upregulation. *World J Gastroenterol* **19,** 6170–7 (2013).

46. Liu, Z. Y. *et al.* Retinoblastoma protein-interacting zinc-finger gene 1 (RIZ1) dysregulation in human malignant meningiomas. *Oncogene* **32,** 1216–1222 (2013).

47. Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B. & Carstens, R. P. ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell* **33,** 591–601 (2009).

48. Ueda, J. *et al.* Epithelial splicing regulatory protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. *Oncogene* **33,** 4485–4495 (2014).

49. Gordon, G. M., Lambert, Q. T., Daniel, K. G. & Reuther, G. W. Transforming JAK1 mutations exhibit differential signalling, FERM domain requirements and growth responses to interferon-γ. *Biochem. J.* **432,** 255–265 (2010).

50. Ren, Y. *et al.* JAK1 truncating mutations in gynecologic cancer define new role of cancer-associated protein tyrosine kinase aberrations. *Sci. Rep.* **3,** (2013).

51. Einav, U. *et al.* Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene* **24,** 6367–6375 (2005).

52. Caspi, E. & Rosin-Arbesfeld, R. A novel functional screen in human cells identifies MOCA as a negative regulator of Wnt signaling. *Mol. Biol. Cell* **19,** 4660–4674 (2008).

53. Taupin, D. *et al.* A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Hum. Genome Var.* **2,** 15013 (2015).

54. Howitt, B. E. *et al.* Association of Polymerase e-Mutated and Microsatellite-Instable Endometrial Cancers With Neoantigen Load, Number of Tumor-Infiltrating Lymphocytes, and Expression of PD-1 and PD-L1. *JAMA Oncol.* **1,** 1319–1323 (2015).

55. Lee, V., Murphy, A., Le, D. T. & Diaz, L. A. Mismatch Repair Deficiency and Response to Immune Checkpoint Blockade. *The Oncologist* **21,** 1200–1211 (2016).

56. Lujan, S. A., Clark, A. B. & Kunkel, T. A. Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects. *Nucleic Acids Res.* **43,** 4067–4074 (2015).

57. Huang, FW. *et al.*, Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

58. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).

59. Rheinbay, E *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60 (2017).

Figure Legends:

Figure 1: Identifying somatic indels in microsatellites (MS indels) – schematic description of MSMuTect**. A.** All reads containing an MS region and sufficient 3' and 5' flanking sequence are aligned to a collection of all MS loci and the number of reads supporting each MS length are tallied to create a histogram of observed read lengths per locus. **B.** The length histograms for all sites that share the same underlying motif and number of repeats (i.e., sites with the same motif and mode length) from the X chromosome of male normal samples were combined into a single histogram. This combined histogram represents the empirical noise distribution (i.e., the probability that a true allele with ***i*** repeats will generate a read with ***j*** repeats). **C.** The maximum likelihood method and empirical noise distribution are used to identify the set of alleles that best describes the histogram for a given locus. This set includes the number of alleles, the length of each allele, and the fraction of DNA molecules representing each allele in the sample. After determining the most likely allele for both the tumor and normal sample, somatic MS indels are nominated when the tumor model fits the tumor data better than the normal model fits the tumor data and vice versa (Online Methods).

Figure 2: Distribution of MS indels across 6,747 tumors from 20 tumor types. Red horizontal lines represent the median fraction of MS indels in each tumor type. **Fig. S6** shows a comparison with the SNV distributions for each tumor type.

Figure 3. Differences in mutation patterns and MS indel characteristics between microsatellite unstable (MSI) and microsatellite stable (MSS) tumors. **A**. Distribution of A motif MS indels across clinical microsatellite (MS) subgroups (MS stable [MSS]; high MS instability [MSI-H];

and low MS instability [MSI-L]) in the three TCGA tumor types for which clinical MSI status was reported (colon adenocarcinoma [COAD], stomach adenocarcinoma [STAD], and uterine corpus endometrial carcinoma [UCEC]). Tumors with ≥15% of SNVs attributed to MS mutations (MSI-SNVs; Online Methods) are plotted in red and tumors with <15% MSI-SNVs are shown in blue. Similarly, tumors with ≥15% SNVs attributed to POLE-mediated mutagenesis (POLE-SNVs) are denoted with an 'x' (Online Methods). **B.** Mean (and standard deviation) relative MS indel frequencies across quintiles of replication times calculated for MSI-H and MSS tumors (combined from the COAD, STAD and UCEC cohorts) . Correlation between MS indel frequency and replication timing – not significant in MSS tumors (slope = -0.03, Pearson correlation = -0.47, *P*=0.43, t-test), weak but significant negative correlation in MSI tumors (slope = -0.1 Pearson correlation = -0.995, *P*<3x10^{-4}, t-test). **C.** MS indel frequency as a function of MS length are shown for MSS and MSI-H tumors. In both MSS and MSI-H tumors, the mutation frequency increases with increasing MS length. The increase is more rapid in MSI-H tumors based on the ratio of mutation frequency of MSI-H to MSS tumors across MS loci lengths (inset). **D**. Log 10 of the frequencies of MS insertions and deletions as a function of normal and mutated repeat number. The estimated number of MS repeats in the normal sample (y-axis) vs the the change in the number of repeats in the tumor (x-axis). The frequency of each specific event (i.e., an insertion or deletion of a given length) is based on the fraction of the total number of covered loci across all samples. MSI-H samples (upper panel), MSS samples (lower panel), and summaried data across all alleles (middle panel). MSI-H samples have more deletions while MSS samples have more insertions (p-value <10^{-31}, χ^2 test). Only MS loci with ≥ 5 repeats in both the normal and mutated samples were included.

Figure 4: Transcriptional effects of the *ESRP1* p.K511fs MS indel mutation. **A.** *ESRP1* expression levels are significantly lower in *ESRP1* mutant (p.K511fs) versus wild type (WT) MSI tumors from the UCEC cohort ($P<1.5x10^{-9}$, Mann-Whitney test). **B.** The ratio of *FGFR2* isoform IIIc to IIIb is significantly higher (p-value $<10^{-7}$, Mann-Whitney test) in *ESRP1* mutant tumors compared to WT tumors. Increased ratio of *FGFR2* isoform IIIc to IIIb is associated with epithelial to mesenchymal transition.

Figure 5: Location of *ACVR2A* MS indel mutations in MSI-H stomach adenocarcinoma (STAD) samples. The MS indel hotspot p.K437fs was identified in 52 of 69 cases (MSMutSig $q=2.4\times10^{-7}$) and had not been previously identified in these samples

Table 1: Significantly Mutated MS Loci

Tumor set	Gene	Protein/genomic change	Mutated samples	Expected mutated samples	p-value	q-value	Most common MS indel*
COAD-MSI	*ACVR2A*	p.K437fs g.chr2:148683686_148683693delA	80% (32/40)	6.25% (2.5/40)	6.4×10^{-9}	3.1×10^{-5}	A_8->A_7 (100%)
COAD-MSI	*RNF43*	p.G659fs g.chr17:56435161_56435167delC	40% (16/40)	4.25% (1.7/40)	6.1×10^{-6}	0.015	C_7->C_6 (100%)
COAD-MSI	*DOCK3*	p.T1850fs g.chr3:51417604_51417610delC	39% (14/36)	4.4% (1.6/36)	2.1×10^{-5}	0.08	C_7->C_6 (86%)
STAD-MSI	*ACVR2A*	p.K437fs g.chr2:148683686_148683693delA	75% (52/69)	4.5% (3.1/69)	2.6×10^{-9}	9.1×10^{-6}	A_8->A_7 (100%)
STAD-MSI	*RNF43*	p.G659fs g.chr17:56435161_56435167delC	35% (24/69)	2.9% (2/69)	1.9×10^{-6}	0.0034	C_7->C_6 (100%)
STAD-MSI	*MSH3*	p.K383fs g.chr5:79970915:79970922delA	41% (28/69)	4.5% (3.1/69)	3.2×10^{-5}	0.037	A_8->A_7 (85%)
STAD-MSI	*PRDM2*	p.K1489fs g.chr1:14108749:14108757delA	48% (33/69)	8.7% (6/69)	8.2×10^{-5}	0.07	A_9->A_8 (93%)
UCEC-MSI	*RNF43*	p.G659fs g.chr17:56435161_56435167delC	23% (36/155)	0.7% (1.2/155)	1.6×10^{-6}	0.016	C_7->C_6 (84%)
UCEC-MSI	*DOCK3*	p.T1850fs g.chr3:51417604_51417610delC	23% (33/145)	1.6% (2.3/145)	3.9×10^{-6}	0.019	C_7->C_6 (81%)
UCEC-MSI	*JAK1*	p.N860fs g.chr1:65306997:65307004delA	21% (33/158)	2.2% (3.5/158)	1.45×10^{-5}	0.05	A_8->A_7 (89%)
UCEC-MSI	*ESRP1*	p.K511fs g.chr8:95686611:95686618delA	20% (31/158)	2.2% (3.5/158)	3×10^{-5}	0.076	A_8->A_7 (94%)
UCEC-MSI	*ACVR2A*	p.K437fs g.chr8:95686611:95686618delA	18% (29/157)	2.2% (3.5/157)	5.9×10^{-5}	0.096	A_8->A_7 (93%)

*the percentage of tumors harboring the most common MS indel is shown in parentheses

Online Methods

Data description

Whole exome sequence (WES) data from 20 tumor types were downloaded from The Cancer Genome Analysis (TCGA) website (https://tcga-data.nci.nih.gov/docs/publications/tcga/) [60] (**Table S2**). We restricted our analysis to fresh frozen samples sequenced on an Illumina platform.

For the analysis of microsatellite stable (MSS) versus unstable (MSI) tumors, only samples from the colon (COAD), stomach (STAD), and uterine (UCEC) cohorts that had MSI status annotated by the TCGA were used (https://tcga-data.nci.nih.gov/docs/publications/tcga/).

For comparison with previously identified mutations, MAF files were downloaded from the Broad Institute's Genome Data Analysis Center (GDAC; http://gdac.broadinstitute.org/), which includes data from samples used in the TCGA marker papers (https://tcga-data.nci.nih.gov/docs/publications). We also analyzed microsatellite (MS) indels in additional TCGA samples that were not part of the TCGA marker papers, but these did not have a curated MAF file for comparison.

The three BAM files for case NA12878 from the 1000 Genome project[18] (http://www.internationalgenome.org/) which were used for the false positive and false negative analysis were deposited at the FireClouad. All three of these samples were sequenced at the Broad Institute.

Microsatellite (MS) definition and identification

Microsatellites (MS) are genomic regions containing multiple copies of a repetitive motif of 1-6 basepairs (bps). While there is no consensus regarding the number of consecutive motifs required to constitute a microsatellite, we define a MS locus as a sequence with at least five successive motifs, regardless of the motif size. We allowed the MS sequence to have impurities, ie, bases that do not follow the exact repeated motif structure. For example, we considered the sequence …GTCAAAAAAAA***C***AAAAAAAAAATCC… as one MS locus with 17 repeats of an A motif, rather than two MS loci, each containing 8 repeats of an A motif. We allowed up to 15% impurity (ie. up to 15% bases that do not match the exact motif), and used the PHOBOS algorithm[16] with default parameters to identify MS with impurities in both the reference genome and WES sequencing reads. Note that we do not suggest that impurities are errors in the reference genome, but rather reflect our looser definition of MSs. We identified 23,677,217 MS loci in the whole genome, 383,515 MS loci in the regions covered by the TCGA whole exome Illumina data, and 145,516 MS loci in the coding regions (as defined by Oncotator[59]). All exonic MS loci are listed in **Supplementary File S1**.

MS-specific alignment

For each normal and tumor sequencing file (ie. BAM file), we used PHOBOS to identify all reads that contained a MS sequence. Following the approach applied in lobSTR[62], for each MS locus, we used the 5' and 3' flanking sequences of the MS to identify reads that support the specific MS. We considered all reads that had at least 10bp flanking the 5' and 3' ends of the MS. (We found that a minimum of 10bp substantially reduces the number of reads that do not match the particular MS.) The alignment procedure was performed in two steps. First, we created, for each MS motif, a library of segments from the human reference genome (hg19) that contained 100 bases from the 5' and 3' ends of each MS locus. Then for each read that contains a MS sequence, we aligned only the non-MS parts of the sequence to the library that contained loci corresponding to same motif that was found in the read (e.g. a read with 7 AGs was aligned against all MS loci with the AG motif). The second step of alignment was then performed using

Bowtie2[63], and only reads that had a single best alignment were included in downstream analyses. MS-specific alignment decreased the number of incorrectly mapped reads by a factor of ~5 (**Fig. S3**).

Noise estimation

Using the MS-specific alignment, we compiled the set of reads that map to each of the MS loci in every sample. For each MS locus, we generated a histogram of MS repeat lengths **(Fig. 1A)**. We hypothesized that not every length represented in the histogram reflects a true allele in the sample, and that the observed numbers of MS repeats in a read that align to a specific MS locus fluctuate around the true value (or values, in the case of a heterozygous site). Some read lengths may be artifacts that were introduced by polymerase stuttering during PCR, sequencing process, or misalignment. The frequency of such sequencing errors varies across MS loci and depends on parameters such as the specific MS motif and the number of repeats.

To predict the true underlying alleles in the tumor and normal samples, we generated an empirical noise model to estimate, $P_{\{j,m\}}^{\text{Noise}}(k,m)$, the probability of observing a read with a length of *k* repeats of motif *m*, given that the true allele in the sample has *j* repeats of *m.* We assumed that all MS loci with the same motif and the same number of repeats have the same noise distribution (and hence can be pooled together to improve the estimated noise model). In addition, we assumed that all normal samples from male donors have only one true allele at all MS loci on the X chromosome, and that the true number of motif repeats corresponds to the observed mode of repeat lengths (ie. the most common number of repeats), while other repeat lengths represent noise. Using this approach, we generated an empirical noise distribution for MS loci with a specific motif and number of repeats. Finally, we smoothed the noise model using python nonparametric regression function (polynomial of 3rd order)

Allele calling

We used the empirical noise model to infer the most likely allele(s) at each MS locus in every sample. We began with the assumption that the sample had only one allele at a given MS locus, and found the most likely repeat length. In practice, we found the repeat length that maximized the log likelihood,

$$ln\big(\mathcal{L}(A|r_i)\big) = \sum_{\{r_i\}} ln\left(P_A^{\text{Noise}}(r_i)\right)$$

where *A* is the underlying allele, ie. the repeat length of motif *m*, $\{r_\text{i}\}$ represents the set of repeat lengths observed in the reads that mapped to the MS locus, and $P_A{}^{\text{Noise}}$ is the empirical noise model for the allele *A*.

Next, we tested a model in which a sample harbors two distinct alleles at a MS locus present at a given ratio. These two alleles could be either germline (ie. inherited from the two parents) or could represent a somatic mutation at a homozygous site. The ratio between the alleles can be 1:1, as in a germline heterozygous site, or, in tumors, the ratio could vary depending on the number of copies of each allele, the purity of the tumor sample, and whether the mutation appears in all cancer cells or only in a subset of the them. We determined the likelihood for two alleles, A_1 and A_2, with fractions *(f ,1-f)*; e.g. a read with 9 repeats of AC

(*r*=9) and proposed alleles $\vec{A}$ = (A_1=6 AC, A_2=8 AC, *f*=0.4). The contribution of read *r* to the

likelihood function is then given by:

$$ln\left(\mathcal{L}(\vec{A}|r)\right) = ln\left(f \cdot P_{A_1}^{\text{Noise}}(r) + (1-f) \cdot P_{A_2}^{\text{Noise}}(r)\right)$$

And based on all reads at the locus, the log likelihood is:

$$ln\left(\mathcal{L}(\vec{A}|\vec{r})\right) = \sum_{\{r_i\}} ln\left(\mathcal{L}(\vec{A}|r_i)\right)$$

As previously, the allele set that had the maximum likelihood was chosen (by optimizing overt A_1, A_2 and f).

We then compared the two models, the one-allele model and the two-allele model, using the log likelihood ratio test (using a χ^2 null distribution), $P^{\chi^2}(D, \Delta f) < 0.05$ where $D = -2 \cdot \ln(\mathcal{L}_1) + 2 \cdot \ln(\mathcal{L}_2)$ and Δf equals 2, as we added two new parameters – the new allele and its fraction. If the χ^2 test gave a p-value > 0.05, we chose the one-allele model. If the χ^2 p-value<0.05, we repeated the test comparing a two-allele model to a three-allele model, and so forth, until we reached a maximum of four alleles. We applied the following restrictions to this process: (1) we analyzed only sites that had at least 10 reads covering them, and (2) we called an allele only if there were at least 5 reads that support it.

Filtering normal loci

Even though normal samples should not have more than two alleles, we allowed the algorithm to continue scanning for more than two alleles in normal samples as a test to detect MS loci associated with increased noise. We did not call somatic MS indels at sites where the normal samples appeared to have >2 alleles or if the read counts were not consistent with a heterozygous site (i.e., binomial test p-value<0.05 with parameter of 0.5).

Mutation calling

For each tumor/normal pair, after inferring the alleles at each MS locus in each sample separately, we compared the inferred alleles in the tumor and normal samples. MS loci that had different alleles in the tumor and normal samples were considered as potentially having somatic mutations, and were nominated for downstream analysis. To ensure that alleles are indeed different, we tested whether the tumor data is described by the tumor alleles better than the normal alleles, and *vice versa*. This was performed by comparing the Akaike information criterion (AIC) score for the two models and requiring that the difference exceeds a predefined threshold, T_r (this was one of the parameters that were later optimized based on the simulated data):

$$AIC^{Tumor\ model}(Tumor\ data) - AIC^{Normal\ model}(Tumor\ data) > T_r$$
$$AIC^{Normal\ model}(Normal\ data) - AIC^{Tumor\ model}(Normal\ data) > T_r$$

Finally, as an additional filter, we performed a Kolmogorov-Smirnov (KS) test between the tumor and normal repeat length histograms. The KS test can identify sites with different alleles but does not identify the exact alleles in the tumor and normal. The KS test p-value was used as another filtering criteria (optimized using the simulated data).

False positive analysis

The false positive (FP) rate was estimated by analyzing three independent whole exome sequencing data sets from sample NA12878 from the 1000 Genomes project (each with an average depth of 60x): NA12878_47, NA12878_49 and NA12878_51. All three of these samples were sequenced at the Broad Institute, each based on a different WES library (to capture the variability introduced by library construction as well as by sequencing). From these three files, we created six tumor-normal pairs by selecting one to represent the tumor and a different one to represent the normal. Note that MSMuTect is not symmetric with respect to the tumor and normal (hence the 6 possible pairs) since the tumor can have more than 2 alleles with different allelic ratios whereas the normal is allowed at most two alleles that are consistent with a 1:1 ratio.

Since all data were acquired from the same sample, all putative somatic MS indels identified by MSMuTect are false positives.

We used MSMuTect to call somatic MS indels across a range of parameter settings and estimated the false positive rate by calculating the average number of apparent somatic MS indels nominated across the six pair-wise comparisons (**Online Methods** and **Fig. S4**). We found that the FP rates of the A motif and the C motif are similar across the range of T_r and *KS* parameters (**Fig. S4).** The AC and AG motifs had only ~2000 loci, and our analysis did not yield any FP mutations for either of these motifs. Therefore, we could not independently estimate the FP rates, but we assumed them to be similar to the FP rates of the A and C motifs and therefore used the same parameter values for all motifs. We chose parameters such that the FP rates for the different motifs resulted in an average of ~5 false positive MS indels across the entire exome, consistent with the FP cutoff used in MuTect [62]. To achieve this, we chose values of AIC *Tr*=8 and KS-test = 0.031 for all the motifs.

True positive analysis

To evaluate sensitivity, we simulated 20,000 somatic MS indels by inserting or deleting a single motif repeat at different loci throughout the exome and then measured the ability of MSMuTect to detect these changes as a function of the original number of motif repeats and the variant allele fraction. We chose to insert or delete a single motif since these are the most prevalent MS indel events in the genome and are also the most challenging to detect.

We first created virtual tumor datasets using the same three WES datasets from sample NA12878 (NA12878_47, NA12878_49 and NA12878_51). Here, we defined NA12878_47 as the normal sample and NA12878_49 as the tumor sample and simulated MS indels using data from NA12878_51. We generated somatic MS indels by replacing a fraction *fr* of read lengths in a histogram representing a site with *k* repeats with read lengths from a site with *l* repeats, thus representing a somatic event from *k* to *k,l* at (1-*fr*,*fr*).

We then used MSMuTect to detect somatic mutations by comparing the simulated tumor to the third copy of NA12878 (acting as the matched normal). We evaluated the sensitivity of MSMuTect to identify MS indels for various allele fractions and repeat lengths (**Fig. S5**). We evaluated MSMuTect using different values of *fr* (ranging from 0.05 to 0.5 with steps of 0.05) and generated 200 mutations for each allele (*k*), mutated at random, to alleles $l=k\pm1$. Sensitivity was highest for shorter MS loci (e.g. sensitivity decreased from 98% for AAAAA, or A_5 in short, to 75% for A_{12}) (**Fig. S5**). Simulated MS indels with an allele frequency below 20% exhibited high rates of false negatives, likely because the allele fraction of ‘artificial’ MS indels generated by PCR exceeded the simulated MS indel fraction.

RNA Validation

For the list of the 7 significant MS loci, we manually compared the 161 MS indels found in the stomach cancer (STAD) cohort to the corresponding tumor RNA-seq data which was obtained from the Broad Institute's Genome Data Analysis Center (GDAC; http://gdac.broadinstitute.org/). An indel was confirmed if at least two RNA-seq reads supported the mutant MS allele (**Table S3**).

MSI and POLE classification

For each sample, a score associated with *POLE* mutations and a score associated with MSI mutations were calculated based on the ratio of *signal* mutations (i.e., mutations uniquely associated with the mutational process) to *background* mutations (other mutations). For *POLE*, the signal mutation[63] is C>A in the context TCT, and the background mutations are all other C>A mutations. The other common *POLE*-associated mutation – C>T in the context TCG – was not used as a signal mutation because it is also present in other common mutational processes, including the signature associated with spontaneous cytosine deamination at meCpG dinucleotides (sometimes called the "aging" signature), and APOBEC-associated signatures[65]. For MSI, a set of three signal mutations were chosen: C(C>A)N, G(C>T)N, and Y(A>G)N (where Y is a pyrimidine and N is any base) based on previous analyses[23], and all other mutations were considered background mutations. Finally, we applied a sigmoid function to the ratio of these mutation counts to produce a final score value between 0 and 1.

Cancer genes

We used a list of 727 widely accepted cancer genes recently published by Nik-Zainal et. Al,[6] (see Supplementary Table 12 therein), which combined genes from the Cancer Gene Census (CGC)[66] list with gene lists from other accepted sources and recent publications.

Diversity in normal samples

As part of MSMuTect, we identify the MS alleles in the normal samples before comparing them to tumor alleles. For each MS locus, we analyzed the alleles across all normal samples and calculated its *diversity*, i.e. the fraction of normal samples that had an allele that is different from the reference genome. For the significance analysis (both for MSMutSig and the search for new events in known cancer genes), we excluded loci that had >10% diversity. While this is similar to the rationale for using a panel-of-normals comparison to exclude sites with either missed germline events or sequencing artifacts[64], in MS loci, this approach may also identify sites that are more prone to MS indels and have a naturally higher mutation rate.

MSMutSig

MSMutSig searches for MS loci that are mutated significantly more frequently than expected by chance. We found that the main two covariates that influence the mutation rate at MS loci are the specific motif and the number of repeats (**Figs. 3B-C**), while other covariates that are known to influence SNV rates (such as replication timing) have minimal effect on MS mutation rates. Thus, we estimated the background mutation frequency for each motif and repeat length in every tumor type separately.

We estimated the rates (and tested the significance) of loci that contained at least one MS mutation across the analyzed cohort. We calculated these conditional rates (ie. conditional on observing at least one event) since we observed a wide variability of mutation rates with a significant enrichment of sites with no mutation. Estimating the mutation rate including these “stable” sites will underestimate the overall background rate and hence inflate the list of significantly mutated loci. As an example, for the A motif with 11 repeats, there were 208/242 loci without any MS indel across the COAD MSI-H cohort, which is ~6 times more than we would have expected (35 loci, P-value$<10^{-16}$ Binomial test,) when using all sites and events to estimate the background rate. Therefore, we concluded that there is a subset of MS loci that are less prone to MS indels and should be excluded from the estimation. Even after excluding these "stable" sites, there was still a large variability of mutation rates among MS loci with the same motif and repeat length, beyond the variability one would expect from a binomial distribution assuming all sites had the same underlying background mutation rate. This high variability was observed even among loci that reside in genomic regions that are less likely to harbor functionally relevant MS loci than exons, such as UTR's and introns (**Fig. S9**).

Therefore, we included an additional variable to attempt to capture this increased variability. We used a negative-binomial distribution (also known as the gamma-Poisson), which has two parameters that control the mean and the variability around the mean. We set the mean to reflect the average mutation rate (at sites with at least one MS indel), and then tuned the variability such that no significant loci were identified outside the exome (with FDR q<0.1). We then used these parameters to identify significantly mutated MS loci in the coding regions. The Q-Q plots for the non-coding MS loci and coding MS loci are shown in Fig**s. S10-S12** (for different tumor types). One can see that there is no inflation of significantly mutated sites and most MS loci follow the expected uniform p-value distribution (ie. reside close to the diagonal of the Q-Q plot).

Expression data

The RNAseq based normalized expression level for each gene was obtained from the Broad Institute's Genome Data Analysis Center website (http://gdac.broadinstitute.org/). We used the $\log_2$-normalized RSEM values when available, but in cases where they were not available, we used $\log_2$ RPKM values.

References

60. The Cancer Genome Atlas - Data Portal. Available at: https://tcga-data.nci.nih.gov/docs/publications/tcga/. (Accessed: 10th October 2016)

61. Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36,** E2423–E2429 (2015).

62. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).

63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

64. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

65. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

66. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).