





Analysis of somatic microsatellite indels identifies driver events in human tumors

Yosef E Maruvka^{1,2}, Kent W Mouw^{3,4}, Rosa Karlic⁵ , Prasanna Parasuraman¹, Atanas Kamburov^{1,2}, Paz Polak^{1,2} , Nicholas J Haradhvala^{1,2}, Julian M Hess², Esther Rheinbay^{1,2}, Yehuda Brody², Amnon Koren⁶, Lior Z Braunstein^{1,2}, Alan D'Andrea²⁻⁴, Michael S Lawrence^{1,2}, Adam Bass^{2,7}, Andre Bernards¹, Franziska Michor^{2,3,8,9} , & Gad Getz^{1-3,10} 

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 has been largely unexplored, owing 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 a higher frequency than expected by chance. Applying MSMuTect to whole-exome data from 6,747 human tumors representing 20 tumor types, we identified >1,000 previously undescribed MS indels in cancer genes. Additionally, we demonstrate that the number and pattern of MS indels can accurately distinguish microsatellite-stable tumors from tumors with microsatellite instability, thus potentially improving classification of clinically relevant subgroups. Finally, we identified seven 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*).

MSs are regions of the genome characterized by repetition of a short sequence motif (usually 1–6 bp)¹. MSs are abundant in nontranscribed regions of the human genome but also occur in exons and untranslated regions (Supplementary Fig. 1). In the germ line, rates of insertions and deletions (indels) in MSs are significantly higher than rates of single-nucleotide substitutions elsewhere in the genome (10^{-4} to 10^{-3} compared with $\sim 10^{-8}$ per locus per generation, respectively)². The increased mutation rate within MS indels is thought to arise because of DNA polymerase slippage during replication, thus leading to changes in the number of repeats. MS indels frequently result in frameshift mutations and can therefore dramatically alter protein function or expression¹.

More than 40 hereditary diseases are caused by germline MS indels^{3–5}. In addition, many cancer-associated genes⁶ (e.g., *PTEN* and *NF1*) contain MS loci, and in some cases, somatic MS indels have been causally implicated in cancer⁷. Tumors with microsatellite instability (MSI) have dramatically higher numbers of MS indels, owing to a loss of normal mismatch repair (MMR) function⁸. 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 importance, somatic MS indels have not been systematically analyzed in cancer, owing to challenges associated with their detection in current massively parallel sequencing data, including read-length limits and PCR errors¹³. The frequency of such sequencing errors varies significantly across MS loci (Online Methods); therefore, methods using 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, because the background mutation frequency can vary by nearly two orders of magnitude¹⁴. 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 calls¹⁵.

To address these challenges, we developed two new tools: MSMuTect (Supplementary Software 1) for detecting somatic MS indels from sequencing data and MSMutSig (Supplementary Software 2) for

¹Massachusetts General Hospital Center for Cancer Research, Charlestown, Massachusetts, USA. ²Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ³Harvard Medical School, Boston, Massachusetts, USA. ⁴Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ⁵Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia. ⁶Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA. ⁷Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ⁸Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ⁹Department of Stem Cells and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. ¹⁰Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to G.G. (gadgetz@broadinstitute.org).

Received 14 November 2016; accepted 18 August 2017; published online 11 September 2017; doi:10.1038/nbt.3966

detecting loci with a significantly higher-than-expected frequency of MS indels. Applying these tools across 6,747 tumors representing 20 tumor types, we uncovered unique properties of MS indels and identified MS loci that probably represent cancer driver events. Comparison of MS indels across clinical MS groups revealed differences between microsatellite stable (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 from whole-exome sequencing data of 6,747 tumor-normal pairs across 20 tumor types from The Cancer Genome Atlas (TCGA) to the unique sequences flanking 383,515 MS loci, defined as sites with at least five repeats of a 1- to 6-bp motif in the exome territory¹ (Online Methods and **Supplementary Fig. 2**). This re-alignment step, compared with the standard alignment, decreased the fraction of misaligned reads (**Supplementary Fig. 3**). For every MS locus, we then counted 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 alleles 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 with 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 represented 98% of the MS loci in the exome (Online Methods).

To accurately identify the alleles present in the tumor and normal samples and to detect somatic MS indels, we used these noise models to calculate, for each MS locus, the set of most likely alleles (**Fig. 1c** and Online Methods). For each locus, we used a log-likelihood ratio test to compare models in which the locus contained 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 with a three-allele model, and so forth, to a maximum of four alleles. Finally, in cases in which the set of alleles was different between the tumor and normal samples, 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** and Online Methods).

We tested the sensitivity and specificity of MSMuTect by using an approach similar to that previously described for MuTect¹⁶. We analyzed sequencing replicates from a single individual and selected parameters that, on average, generated no more than five false positives per exome (**Supplementary Fig. 4** and Online Methods). To evaluate sensitivity, we simulated MS indels by inserting or deleting a single motif repeat at different loci throughout the genome.

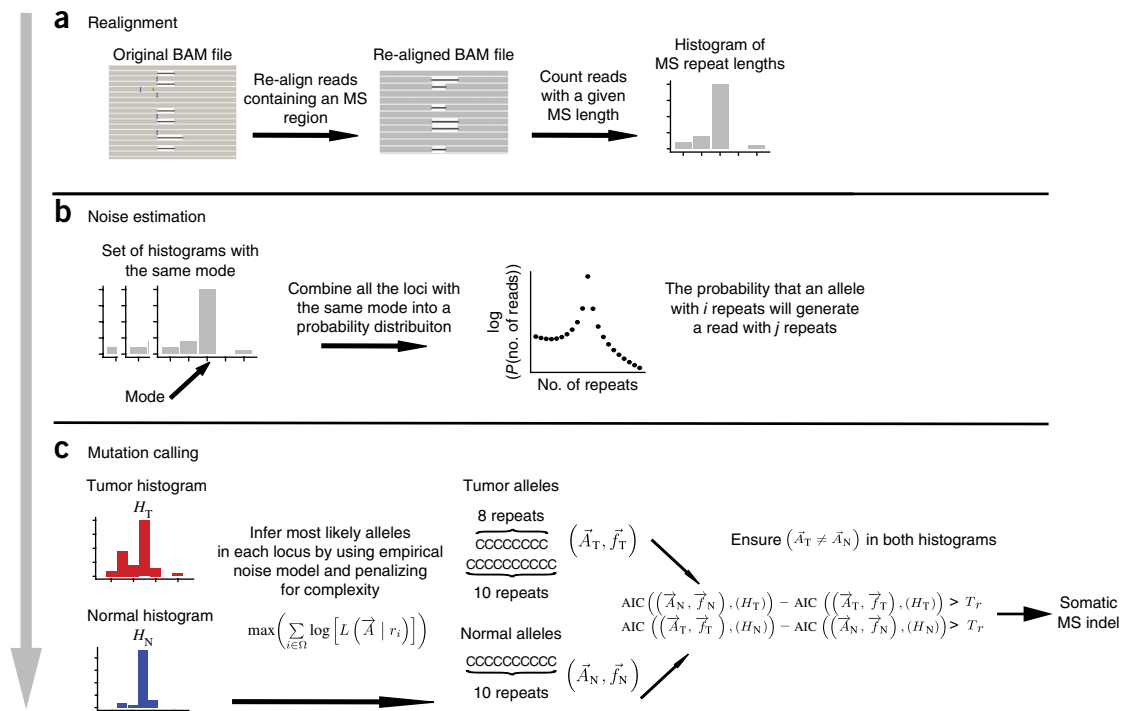


Figure 1 Identifying somatic indels in microsatellites (MS indels): schematic description of MSMuTect. **(a)** All reads containing an MS region and sufficient 3' and 5' flanking sequence are aligned to a collection of all MS loci, and the number of reads supporting each MS length are tallied to create a histogram of observed repeat lengths per locus. **(b)** The repeat-length histograms for all sites sharing 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 the most likely allele is determined for both the tumor (T) and normal (N) samples, 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). Variables are defined in Online Methods.

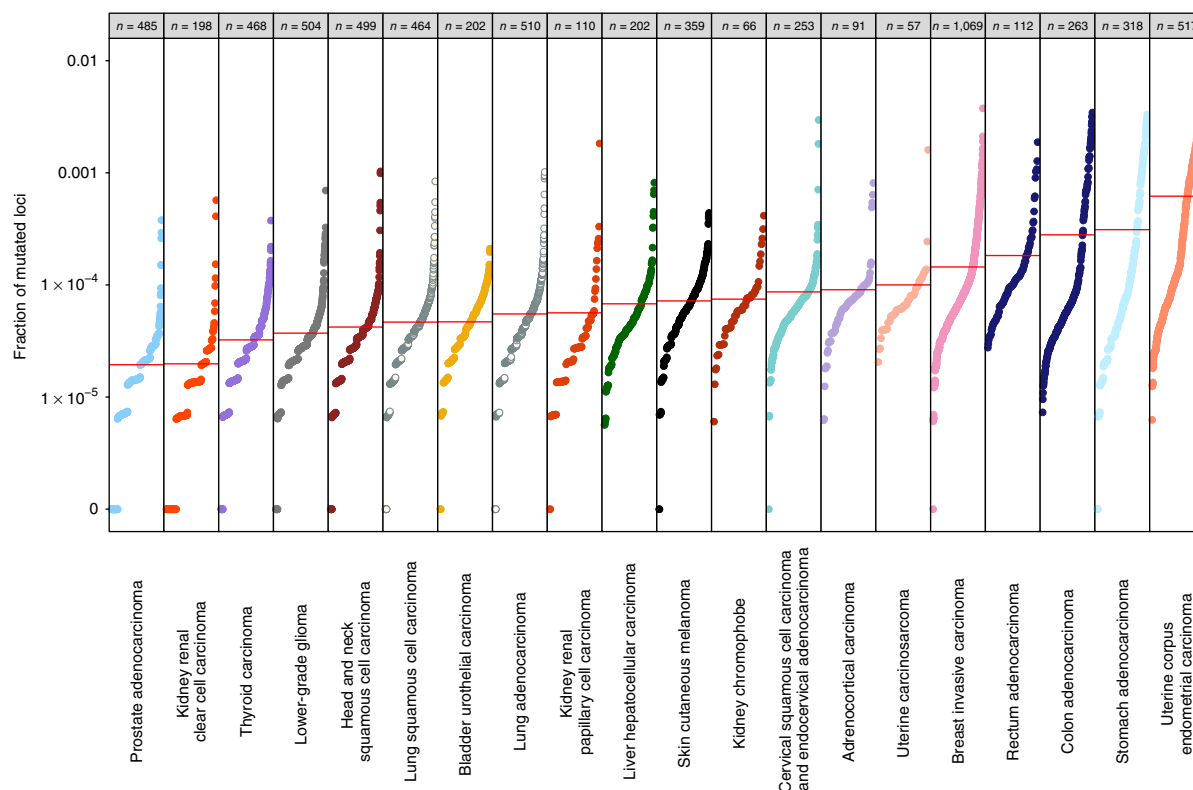


Figure 2 Distribution of MS indels across 6,747 tumors from 20 tumor types. Red horizontal lines represent the mean fraction of mutated MS loci in each tumor type. **Supplementary Figure 6** shows a comparison with the SNV distributions for each tumor type.

We evaluated the sensitivity across various allele fractions and MS locus lengths and found that it was higher for shorter MS loci and decreased when the MS indel allele fraction fell below 20% (**Supplementary Fig. 5** and Online Methods).

MS indel mutational landscape

We applied MSMuTect across 6,747 TCGA whole-exome tumor-normal pairs representing 20 tumor types (**Supplementary Tables 1** and **2**). Our analysis identified 174,638 MS indels, with a range of 0 to 900 per tumor. We observed extensive inter- and intratumor variability in the MS indel frequency a result similar to the variability reported for single-nucleotide variations (SNVs) and copy-number alterations¹⁵ (**Fig. 2** and **Supplementary Fig. 6**). The average MS indel frequency varied significantly across tumor types, and the highest frequencies were found in colorectal (COAD and READ), stomach (STAD), and endometrial tumors (UCEC), in agreement with the MMR deficiency frequently observed 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 features¹⁷. Moreover, a recent study⁶ has identified mutational signatures consistent with loss of mismatch repair in BRCA. In a recent report, Hause *et al.*¹³ did not identify MSI-H cases in TCGA BRCA; however, the authors have analyzed a subset of the TCGA breast cancer cohort (266/1,069 cases) including only 3/36 tumors in which we identified >100 MS indels. Previous reports^{18–20} have identified a small fraction of MSI cases in cohorts of cervical squamous cell carcinoma and endocervical adenocarcinoma, uterine carcinosarcoma, and adrenocortical carcinoma, and our analysis indeed identified MS indels in these tumor types (ranked sixth, seventh and eighth in average MS indel frequency, respectively).

To validate the identified MS indels, we analyzed RNA-seq data available for a subset of the samples (**Supplementary Table 3**). For each of the 150 significantly mutated MS indels (described below) with sufficient RNA-seq coverage (four or more reads), we manually compared the alleles inferred by MSMuTect with the alleles observed in the RNA-seq data and validated 87% of them (**Supplementary Table 3**). Importantly, RNA-seq probably 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%); **Supplementary Table 3**). 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 distinct alternate alleles).

MSMuTect correctly classifies tumors with respect to MS stability

We next asked whether MSMuTect could recapitulate independent measures of tumor MS stability. Tumors from the COAD, STAD, and UCEC cohorts, as part of TCGA, 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 two or more MS loci), by using a PCR-based assay to assess size variability at the five Bethesda MS loci¹². Applying MSMuTect, we found that tumors classified as MSI-H had significantly more MS indels than did samples that were MSS or MSI-L (MSI-H versus MSS: COAD, median 104.5 versus 3.0, $P < 10^{-22}$; STAD, 64.5 versus 1.0, $P < 10^{-28}$; UCEC, 94.5 versus 2.0, $P < 10^{-58}$,

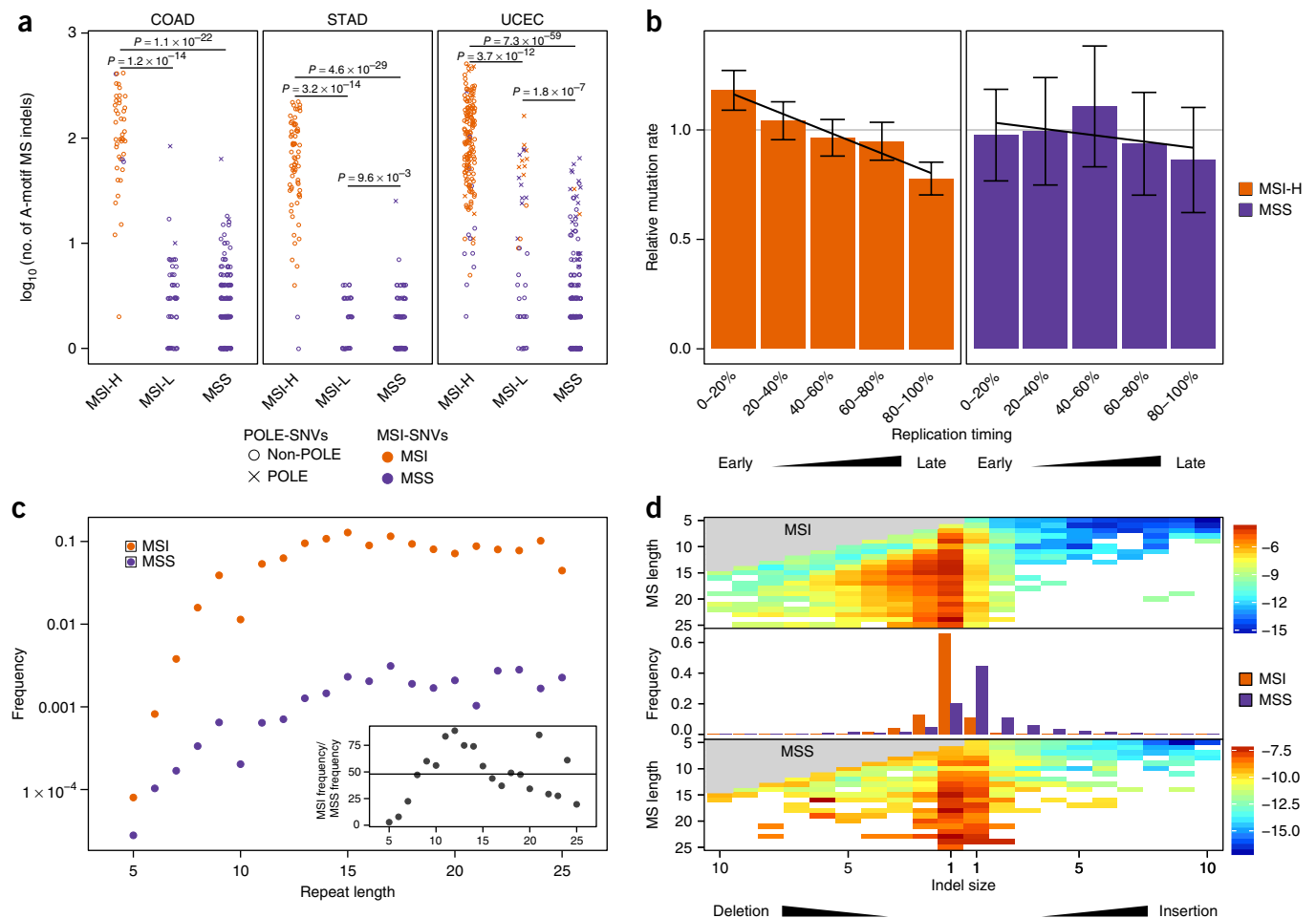


Figure 3 Differences in mutation patterns and MS indel characteristics between MSI and MSS tumors. **(a)** Distribution of A-motif MS indels across clinical MS subgroups (MSS, MSI-H, and MSI-L) in the three TCGA tumor types for which clinical MSI status was reported (COAD, STAD, and UCEC). Tumors with $\geq 15\%$ of SNVs attributed to MSI mutations (MSI-SNVs; Online Methods) are plotted in red, and tumors with $< 15\%$ MSI-SNVs are shown in blue. Similarly, tumors with $\geq 15\%$ SNVs attributed to POLE-mediated mutagenesis (POLE-SNVs) are denoted with an 'x' (Online Methods). **(b)** Mean (\pm s.d.) relative MS indel frequencies across quintiles of replication times calculated for MSI-H and MSS tumors (combined from the COAD, STAD, and UCEC cohorts). The correlation between MS indel frequency and replication timing was not significant in MSS tumors (slope = -0.03 , Pearson correlation = -0.47 , $P = 0.43$, two-tailed t -test) but showed a weak but significant negative correlation in MSI tumors (slope = -0.1 , Pearson correlation = -0.995 , $P < 3 \times 10^{-4}$, two-tailed t -test). **(c)** MS indel frequencies as a function of MS length are shown for MSS and MSI-H tumors. In both MSS and MSI-H tumors, the mutation frequency increased with increasing MS length. The increase was more rapid in MSI-H tumors, on the basis of the ratio of the mutation frequency of MSI-H to MSS tumors across MS locus lengths (inset). **(d)** log₁₀ of the frequencies of MS insertions and deletions as a function of normal MS repeat length and tumor indel size. The estimated number of MS repeats in the normal sample (y axis) versus 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 (top), MSS samples (bottom), and summarized data across all alleles (middle). MSI-H samples had more deletions, whereas MSS samples had more insertions ($P < 10^{-31}$, χ^2 test). Only MS loci with five or more repeats in both the normal and mutated samples were included.

two-tailed Mann–Whitney; **Fig. 3a** and 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 for UCEC and STAD (UCEC, median of 9 versus 2, $P < 10^{-6}$; STAD, 3 versus 2, $P < 10^{-3}$, one-tailed Mann–Whitney), because of the 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 ($P < 10^{-16}$, one-tailed t -test, **Supplementary Fig. 7**) more likely to have several MS indels at the same locus, at both germline homozygous and heterozygous MS loci.

Although MSMuTect was able to separate most 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 phenotype²¹ (Online Methods). As expected, tumors in which MSI-SNVs composed $> 15\%$ 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 (fewer than ten) had an MSI-SNV fraction $< 15\%$ (blue in **Fig. 3a**), thus 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 numbers of MS indels also had a relatively high number of total SNVs (**Fig. 3a**). Mutations in the exonuclease (proofreading) domain of DNA polymerase ϵ (encoded by *POLE*) can dramatically

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 ^a
COAD-MSI	<i>ACVR2A</i>	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	<i>RNF43</i>	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	<i>DOCK3</i>	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	<i>ACVR2A</i>	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	<i>RNF43</i>	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	<i>MSH3</i>	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	<i>PRDM2</i>	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	<i>RNF43</i>	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	<i>DOCK3</i>	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	<i>JAK1</i>	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	<i>ESRP1</i>	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	<i>ACVR2A</i>	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%)

^aThe percentage of tumors bearing the most common MS indel is shown in parentheses.

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 probably 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 δ (encoded by *POLD1*; $n = 2$). Although most of the POLE/*POLD1*-mutated tumors (45/63) were classified as MSS or MSI-L, they had significantly more MS indels than did the other MSS and MSI-L tumors (median 54 versus 2 among MSI-L and 18.5 versus 2 among MSS), thus raising the possibility that POLE/*POLD1* exonuclease-domain mutations may contribute to the MS indel burden and also 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 the numbers of MS indels, MSI and MSS samples also differed in their association between MS indel frequency and DNA-replication timing, and both exhibited associations distinct from that reported for SNVs^{14,22}. In general, unlike SNV density, MS indel density did not show a strong correlation with replication timing. Interestingly, there was no correlation in MSS samples (slope = -0.03, Pearson correlation = -0.47, $P = 0.43$, two-tailed t -test; Fig. 3b), and there was a marginal but significant decrease with replication timing in MSI samples²³ (slope = -0.1, Pearson correlation = -0.995, $P = 0.0003$, two-tailed t -test; Fig. 3b), a result opposite from the direction observed for SNVs.

Likewise, in both MSI and MSS tumors, MS indels were more common at loci with longer repeat lengths; however, the slope and shape of these relationships differed (Fig. 3c). Moreover, the ratio of insertions to deletions was different between MSS and MSI cases, and MSI cases had a tendency toward deletions²³, whereas MSS cases tended toward insertions (Fig. 3d; $P < 10^{-31}$, χ^2 test). The tendency to increase repeat lengths in MSS cases was consistent with germline MS indels, which have been shown to preferentially undergo insertions in MS loci with <15 repeats (ref. 2).

MS indels in known cancer genes

We next sought to identify somatic MS indels that drive tumorigenesis. We first attempted to identify previously undescribed MS indels across 727 known cancer genes⁶ in a cohort of 4,064 TCGA samples with curated mutation calls (Online Methods). We focused our analysis on MS loci for which the inferred germline allele matched the reference genome in at least 90% of the normal (germline) samples (Supplementary Fig. 8). MS indels at loci with greater germline diversity may have

weaker functional effects or represent noisy sites. We detected 1,470 MS indels across these genes (Supplementary Table 4), including 89 indels that had been previously identified by the TCGA consortium and 1,105 indels in samples without any other indel or nonsynonymous 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 those cases, the identified MS indel may represent the 'second hit'²⁴. In some genes, previously unidentified MS indels comprised a substantial fraction of the total number of mutations. Reassuringly, we found that MS indels were enriched in tumor-suppressor genes²⁵ compared with oncogenes (993 MS indels in 70 tumor-suppressor genes versus 272 MS indels in 53 oncogenes, as well $P < 10^{-58}$, one-tailed binomial test).

MSMutSig, a tool for identifying driver MS indels

Next, we extended our MutSig suite of tools¹⁵ 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 and 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 contained more (or fewer) mutations than predicted by the model (Supplementary Fig. 9). To address this discrepancy, we applied a more dispersed distribution, the negative binomial, and fit the extra dispersion parameter such that no MS loci in noncoding regions would be nominated as significantly mutated (at Benjamini-Hochberg false discovery rate $q < 0.1$). This model indeed captured the variability of MS indel rates in coding regions as well (Supplementary Figs. 10–12).

After it had been 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 three significant MS loci in the genes *ACVR2A*, *RNF43*, and *DOCK3*; in STAD, four loci in *ACVR2A*, *RNF43*, *MSH3*, and *PRDM2*; and in UCEC, five loci in *RNF43*, *DOCK3*, *JAK1*, *ESRP1*, and *ACVR2A*. Thus, our analysis nominated a total of seven MS hotspots in seven genes (Table 1). Three of these genes (*ACVR2A*, *RNF43*, and *JAK1*) have previously been identified as cancer genes on the basis of an increased mutation frequency in one or more tumor types⁶ (Fig. 5). In the TCGA colon

cancer study²⁶, *MSH3* was not nominated as significantly mutated but was noted to be highly mutated on the basis of manual examination of the sequence data. Notably, owing to the high mutability of MS loci, beyond major cancer genome studies, the literature is inconclusive regarding which of the 17,398 genes with MS loci are associated with cancer²⁷. The remaining three genes (*ESRP1*, *PRDM2*, and *DOCK3*) have not been previously identified as cancer-associated genes (discussed below). Previously identified cancer drivers such as *TGFRB2* (ref. 28) and *RPL22* (ref. 29) were absent from our list because their MS loci were excluded from the analysis, owing to high variability in germline samples.

All seven of the significantly mutated MS indels caused a frameshift mutation within an exon. Frameshift mutations typically result in decreased gene expression, because the altered mRNA undergoes nonsense-mediated decay³⁰. However, if a frameshift mutation occurs near the 3' end of a gene, nonsense-mediated decay is less likely to occur³¹. Of the seven MS indels identified here, four (in *ESRP1*, *MSH3*, *JAK1*, and *PRDM2*) led to a significant decrease in mRNA expression levels (Table 1 and Fig. 4), whereas the MS indels in *ACVR2A* and *DOCK3* occurred near the 3' end of the gene and thus were not expected to lead to nonsense-mediated decay. The MS indel in *RNF43* was in the second-to-last exon; however, the presence of this indel did not correlate with a decreased *RNF43* expression level (one-tailed Mann–Whitney $P = 0.4$) and may represent an exception to the '50-bp rule', similarly to observations regarding *UPF1* (refs. 32,33).

Genes nominated by MSMutSig are candidate cancer drivers

The *ACVR2A* gene, encoding activin A receptor type IIA, contains the most frequently mutated MS locus in our list (p.K437fs). We observed *ACVR2A* 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 and 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. In support of a tumor-suppressor role, two studies^{34,35} have shown that expression of wild-type *ACVR2A* in MSI colon cancer cell lines with mutated *ACVR2A* leads to decreased cell growth. When these previously undescribed MS indel events were considered with other reported alterations, *ACVR2A* was among the most frequently mutated genes in colorectal cancer, and mutations were observed in ~20% of all cases.

The gene encoding ring-finger protein 43 (*RNF43*) contained 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 proliferation³⁶. This gene has been reported to be significant in STAD by TCGA³⁷. Giannakis *et al.*⁷ have recently detected the same *RNF43* MS indel through manual review of *RNF43* sequence data and have determined that it is frequently present in colon and endometrial tumors.

The *MSH3* gene, encoding the protein MutS homolog 3, is a member of the MMR pathway, and germline mutations in *MSH3* are known to increase the risk of developing MSI tumors³⁸. 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 formation³⁹.

PRDM2, the gene encoding the protein PR domain 2, is a histone H3 Lys9 methyltransferase that has been implicated as a tumor suppressor in several tumor types⁴⁰. Decreased *PRDM2* expression has been associated with renal cell carcinoma⁴¹, esophageal squamous cell carcinoma⁴², and meningiomas⁴³. We identified the MS indel

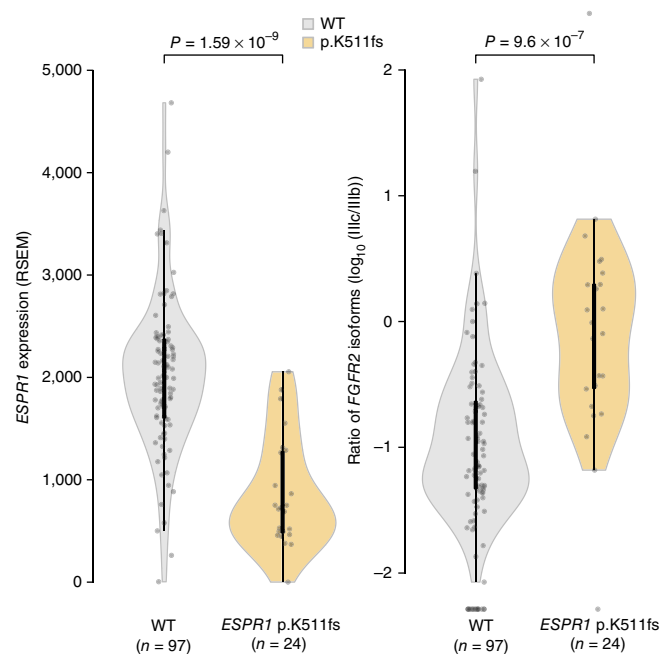


Figure 4 Transcriptional effects of the *ESRP1* p.K511fs MS indel mutation. (a) *ESRP1* transcript levels were significantly lower in the *ESRP1*-mutant (p.K511fs) than the wild-type (WT) MSI tumors from the UCEC cohort ($P < 1.5 \times 10^{-9}$, two-tailed Mann–Whitney test). (b) The ratio of *FGFR2* isoform IIIc to IIb was significantly higher ($P < 10^{-7}$, two-tailed Mann–Whitney test) in *ESRP1*-mutant tumors than WT tumors. The increased ratio of *FGFR2* isoform IIIc to IIb was associated with epithelial–mesenchymal transition.

hotspot p.K1489fs in 48% (33/69) of stomach tumors. Analysis of gene expression data revealed decreased expression in mutated cases ($P = 0.016$, one-tailed Mann–Whitney; Supplementary Fig. 13), a result consistent with partial nonsense-mediated decay.

The epithelial splicing regulatory protein 1 (encoded by *ESRP1*) is a splicing regulator in epithelial cells⁴⁴. Its MS indel hotspot (*ESRP1* p.K511fs) is mutated in approximately 20% (31/158) of MSI endometrial tumors. *ESRP1* regulates alternative splicing of *FGFR2* from the IIIc mesenchymal isoform to the IIb epithelial isoform⁴⁴. Thus, mutations in *ESRP1* may contribute to the epithelial–mesenchymal transition. In pancreatic cancer⁴⁵, the transition from expression of the *FGFR2*-IIb 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; $P < 1.5 \times 10^{-9}$, one-tailed Mann–Whitney) and a significant increase in the ratio of isoform IIIc to IIb in *ESRP1*-mutant cases (Fig. 4b; $P < 9 \times 10^{-7}$, one-tailed Mann–Whitney).

Our finding that *JAK1* contained the frameshift mutation p.N860fs in 21% (33/158) of endometrial tumors (Table 1) was somewhat unexpected, given *JAK1*'s known role as an oncogene⁴⁶. Kim *et al.*²³ have found that the *JAK1* p.N860fs indel is associated with repression of transcript levels of *JAK1* downstream targets, and a recent study⁴⁷ has suggested that truncated *JAK1* modulates the IFN γ signaling pathway and enables tumor immune evasion. We compared expression of an IFN γ -related genes⁴⁸ in tumors with or without the *JAK1* p.N860fs indel and found a significant decrease in expression in 21 of 27 IFN γ -related genes in tumors with the p.N860fs indel. Therefore, *JAK1* loss may promote tumor survival by inhibiting an IFN γ -mediated antitumor immune response.

Finally, *DOCK3*, which encodes the protein dedicator of cytokinesis 3, carried 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 has recently been implicated as an inhibitor of the WNT signaling pathway⁴⁹. *CTNNB1*, encoding a core member of the WNT pathway, was mutated in approximately 30% of endometrial tumors, and *DOCK3* mutations were mutually exclusive to *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 re-alignment 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 decrease 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 number of motif repeats different from that observed in 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 with <10 with MSMuTect, and only an approximately threefold difference in the number of MS indels between MSS and MSI cases (897 in MSS versus 3,009 in MSI) versus an ~18-fold difference with MSMuTect (8 versus 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 alleles than by the normal alleles. A recent report by Kim *et al.*²³ has used the Kolmogorov–Smirnov (KS) test to compare repeat-length distributions in the tumor and normal samples 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 decreasing 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, thus raising the possibility of an inherited pathogenic *RNF43* MS indel. A recent study by Taupin *et al.*⁵⁰ has shown 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 data set 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, thus suggesting that the MSI phenotype also occurs in these tumor types and that MSI testing

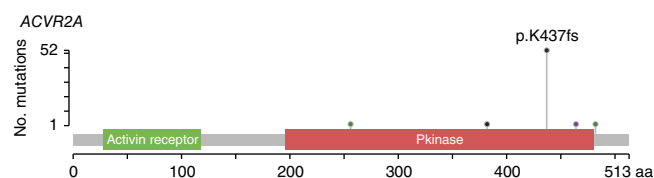


Figure 5 Location of *ACVR2A* MS indel mutations in MSI-H STAD samples. The MS indel hotspot p.K437fs was identified in 52 of 69 cases (MSMutSig $q = 2.4 \times 10^{-7}$) and had not been previously identified in these samples.

may be warranted for these tumors in certain clinical settings, such as when screening for immunotherapy trials^{10,51,52}.

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 the MS indel frequency within each MS subgroup, particularly among endometrial tumors.

We found that many MSI-L tumors had MS indel frequencies similar to those of MSS tumors, thus suggesting that some were misclassified by the MSI assay. However, a subset of MSI-L tumors, particularly endometrial tumors, had MS indel frequencies that more closely resembled those of MSI-H tumors. Proofreading deficient tumors arising from *POLE*/*POLD1* exonuclease mutations had dramatically higher 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 appeared to have concomitant *POLE* mutations and MSI, other MSI-L endometrial tumors had 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 has recently been reported in yeast⁵³. It is possible that a dramatic increase in SNVs resulting from loss of *POLE* proofreading might saturate the capacity for MMR (which corrects both SNVs and indels) and thus indirectly result in a greater number of unrepaired MS indels.

Finally, we observed that many of the MSI-H/MSI-L endometrial tumors with the lowest MS indel frequencies lacked the MSI-SNV signature, thus suggesting that these tumors may have been misclassified as MSI and 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^{10,51}.

On the basis of 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. MutSig¹⁵ has been 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, application of MutSig to the MSI-H endometrial cohort by using all mutations (SNVs, general indels and MS indels) yielded 296 significant genes ($q < 0.1$) and an inflated quantile–quantile plot, thus suggesting an inadequate null model (Supplementary Fig. 14).

On the basis of a high frequency of MS indels, many genes have been proposed to drive cancer²⁷; however, our analysis suggested that many of these MS loci have a high background mutation rate and therefore may be frequently mutated but may not have contributed to

positive selection. Applying MSMutSig across 6,747 cases, we identified seven significantly mutated MS loci, three of which occurred in genes not previously nominated as cancer-associated 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 noncoding regions were not under selective pressure and thus could serve as an upper estimate of the indel mutation rate arising from technical factors (such as PCR errors). However, cancer driver mutations are known to exist in regulatory regions^{54–56}, 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 noncoding 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 somatic events such as SNVs or copy-number alterations should provide a more comprehensive view of cancer driver events.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Mayer for supplying and supporting the PHOBOS tool. G.G. 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.E.M., P. Polak, and A. Kamburov were funded by G.G.'s start-up funds at Massachusetts General Hospital. K.W.M. 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.M. gratefully acknowledges support from the Dana-Farber Cancer Institute Physical Sciences Oncology Center (NIH U54CA193461). R.K. was supported by the European Commission Seventh Framework Programme (Integra-Life; grant 315997) and the Croatian Science Foundation (grant IP-2014-09-6400).

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.K., P. Parasuraman, A. Kamburov, P. Polak, N.J.H., J.M.H., E.R., Y.B., A. Koren, L.Z.B., A.D.A., M.S.L., A.J.B., A.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.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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., Nichol Edamura, K. & 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.* **36**, 115–127 (2017).
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. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
17. Tokunaga, E. *et al.* Frequency of microsatellite instability in breast cancer determined by high-resolution fluorescent microsatellite analysis. *Oncology* **59**, 44–49 (2000).
18. Larson, A.A. *et al.* Analysis of replication error (RER+) phenotypes in cervical carcinoma. *Cancer Res.* **56**, 1426–1431 (1996).
19. Taylor, N.P. *et al.* Defective DNA mismatch repair and XRCC2 mutation in uterine carcinosarcomas. *Gynecol. Oncol.* **100**, 107–110 (2006).
20. Medina-Arana, V. *et al.* Adrenocortical carcinoma, an unusual extracolonic tumor associated with Lynch II syndrome. *Fam. Cancer* **10**, 265–271 (2011).
21. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
22. 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).
23. 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).
24. Knudson, A.G. Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* **68**, 820–823 (1971).
25. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
26. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
27. Cederquist, K. *Genetic and epidemiological studies of hereditary colorectal cancer* PhD thesis, Umeå University (2005).
28. Biswas, S. *et al.* Mutational inactivation of TGFB2 in microsatellite unstable colon cancer arises from the cooperation of genomic instability and the clonal outgrowth of transforming growth factor β resistant cells. *Genes Chromosom. Cancer* **47**, 95–106 (2008).
29. Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
30. Maquat, L.E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5**, 89–99 (2004).
31. 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. USA* **100**, 189–192 (2003).
32. 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* **4**, 801–815 (1998).
33. 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).
34. 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).
35. Ballikaya, S. *Activin receptor type 2 A (ACVR2A)-dependent proteomic and glycomic alterations in a microsatellite unstable (MSI) colorectal cancer cell line model system* PhD thesis, Ruperto-Carola University of Heidelberg (2014).

36. Niu, L. *et al.* RNF43 inhibits cancer cell proliferation and could be a potential prognostic factor for human gastric carcinoma. *Cell. Physiol. Biochem.* **36**, 1835–1846 (2015).
37. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
38. Durauto, 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).
39. de Wind, N. *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).
40. 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).
41. 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).
42. Dong, S.-W. *et al.* Alteration in gene expression profile and oncogenicity of esophageal squamous cell carcinoma by RIZ1 upregulation. *World J. Gastroenterol.* **19**, 6170–6177 (2013).
43. Liu, Z.Y. *et al.* Retinoblastoma protein-interacting zinc-finger gene 1 (RIZ1) dysregulation in human malignant meningiomas. *Oncogene* **32**, 1216–1222 (2013).
44. 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).
45. 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).
46. 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).
47. Ren, Y. *et al.* JAK1 truncating mutations in gynecologic cancer define new role of cancer-associated protein tyrosine kinase aberrations. *Sci. Rep.* **3**, 3042 (2013).
48. 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).
49. 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).
50. Taupin, D. *et al.* A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Hum. Genome Var.* **2**, 15013 (2015).
51. Howitt, B.E. *et al.* Association of polymerase ϵ -mutated and microsatellite-unstable endometrial cancers with neoantigen load, number of tumor-infiltrating lymphocytes, and expression of PD-1 and PD-L1. *JAMA Oncol.* **1**, 1319–1323 (2015).
52. Lee, V., Murphy, A., Le, D.T. & Diaz, L.A. Jr. Mismatch repair deficiency and response to immune checkpoint blockade. *Oncologist* **21**, 1200–1211 (2016).
53. 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).
54. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
55. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
56. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).

ONLINE METHODS

Data description. Whole exome sequencing (WES) data from 20 tumor types were downloaded from TCGA⁵⁷ (**Supplementary Table 2**). We restricted our analysis to fresh-frozen samples sequenced on an Illumina platform.

For the analysis of MSS versus MSI tumors, only samples from the colon (COAD), stomach (STAD), and uterine (UCEC) cohorts that had MSI status annotated by the TCGA were used⁵⁷.

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 MS indels in additional TCGA samples that were not part of the TCGA marker papers, but those did not have a curated MAF file for comparison.

The three BAM files for case NA12878 from the 1000 Genomes Project⁵⁸ (<http://www.internationalgenome.org/>), which were used for the false-positive and false-negative analysis, were uploaded to the Broad Institute's FireCloud platform. All three of these samples were sequenced at the Broad Institute.

Microsatellite definition and identification. MSs are genomic regions containing multiple copies of a repetitive motif of 1–6 bp. Although there is no consensus regarding the number of consecutive motifs required to constitute a microsatellite, we defined an MS locus as a sequence with at least five successive motifs, regardless of the motif size. We allowed the MS sequence to have impurities, i.e., bases that did not follow the exact repeated motif structure. For example, we considered the sequence GTCAAAAAAAAACAAAAAAAAAAATCC (in which the base not matching the motif is in bold italic format) as one MS locus with 17 repeats of an A motif rather than two MS loci, each containing eight repeats of an A motif. We allowed up to 15% impurity (i.e., up to 15% bases that did not match the exact motif), and used the PHOBOS algorithm⁵⁹ with default parameters to identify MSs with impurities in both the reference genome and WES sequencing reads. However, we do not suggest that impurities are errors in the reference genome but instead suggest that they 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 Oncotator⁶⁰). All coding MS indels are listed in **Supplementary Tables 5**.

MS-specific alignment. For each normal and tumor sequencing file (i.e., BAM file), we used PHOBOS to identify all reads that contained an MS sequence. Following the approach applied in lobSTR⁶¹, for each MS locus, we used the 5' and 3' flanking sequences of the MS to identify reads that supported the specific MS. We considered all reads that had at least 10 bp flanking the 5' and 3' ends of the MS. (We found that a minimum of 10 bp substantially decreased the number of reads that did 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 contained an 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 seven AGs was aligned against all MS loci with the AG motif). The second step of alignment was then performed with Bowtie2 (ref. 62), 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 (**Supplementary Fig. 3**).

Noise estimation. Using the MS-specific alignment, we compiled the set of reads that mapped 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 reflected a true allele in the sample and that the observed numbers of MS repeats in a read that aligned to a specific MS locus fluctuated 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 or sequencing, or by misalignment. The frequency of such sequencing errors varied across MS loci and depended 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 had the same noise distribution (and hence could be pooled to improve the estimated noise model). In addition, we assumed that all normal samples from male donors had only one true allele at all MS loci on the X chromosome and that the true number of motif repeats corresponded to the observed mode of repeat lengths (i.e., the most common number of repeats), whereas other repeat lengths represented 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 by using a nonparametric regression function (third-order polynomial) in Python.

Allele calling. We used the empirical noise model to infer the most likely alleles 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,

$$\ln(\mathcal{L}(A | r_i)) = \sum_{\{r_i\}} \ln(P_A^{\text{noise}}(r_i))$$

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

Next, we tested a model in which a sample contains two distinct alleles at an MS locus present in an unknown ratio. These two alleles could either be germline alleles (i.e., inherited from the two parents) or represent a somatic mutation at a homozygous site. The ratio between the alleles could 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 appeared in all cancer cells or only in a subset of cells. We determined the likelihood for two alleles, A_1 and A_2 , with fractions $(f, 1-f)$; e.g., a read with nine AC repeats ($r=9$) and proposed alleles $\vec{A} = (A_1 = 6 \text{ AC}, A_2 = 8 \text{ AC}, f=0.4)$. The contribution of read r to the likelihood function is then given by:

$$\ln(\mathcal{L}(\vec{A} | r)) = \ln(f \times P_{A_1}^{\text{noise}}(r) + (1-f) \times P_{A_2}^{\text{noise}}(r))$$

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

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

As before, the allele set that had the maximum likelihood was chosen (by optimizing over A_1 , A_2 and f). We then compared the two models—the one-allele model and the two-allele model—by using the log likelihood ratio test (with a χ^2 null distribution), $P\chi^2(D, \Delta f) < 0.05$, where $D = -2 \ln(\mathcal{L}_1) + 2 \ln(\mathcal{L}_2)$ and $\Delta f = 2$, because we added two new parameters: the new allele and its fraction. If the χ^2 test yielded a P value > 0.05 , we chose the one-allele model. If the $\chi^2 P$ value was < 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: (i) we analyzed only sites that had at least ten reads covering them, and (ii) we called an allele only if there were at least five reads supporting 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 more than two 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 separately inferring the alleles at each MS locus in each sample, 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 were indeed different, we tested whether the tumor data were described by the tumor alleles better than the normal alleles, and vice versa. This procedure was performed by comparing the Akaike information criterion (AIC) score for the two models and requiring that the difference exceed a predefined threshold, T_r (this was one of the parameters that were later optimized on the basis of the simulated data):

$$\begin{aligned} \text{AIC}^{\text{normal model}}(\text{tumor data}) - \text{AIC}^{\text{tumor model}}(\text{tumor data}) &> T_r \\ \text{AIC}^{\text{tumor model}}(\text{normal data}) - \text{AIC}^{\text{normal model}}(\text{normal data}) &> T_r \end{aligned}$$

Finally, as an additional filter, we performed a KS test between the tumor and normal-repeat-length histograms. The KS test can identify sites with different alleles but cannot identify the exact alleles in the tumor and normal samples. The KS-test P value was used as another filtering criterion (optimized by 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 60×): NA12878_47, NA12878_49, and NA12878_51. All three of these samples were sequenced at the Broad Institute, each on the basis of 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. Notably, MSMuTect is not symmetric with respect to the tumor and normal (hence the six possible pairs), because the tumor can have more than two alleles with different allelic ratios, whereas the normal sample is allowed at most two alleles that are consistent with a 1:1 ratio. Because all data were acquired from the same sample, all putative somatic MS indels identified by MSMuTect were false positives.

We used MSMuTect to call somatic MS indels across a range of parameter settings and estimated the FP rate by calculating the average number of apparent somatic MS indels nominated across the six pairwise comparisons (Online Methods and **Supplementary Fig. 4**). We found that the FP rates of the A motif and the C motif were similar across the range of T_r and KS parameters (**Supplementary Fig. 4**). The AC and AG motifs had only ~2,000 loci, and our analysis did not yield any FP mutations for either of these motifs. Therefore, we were not able to 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 approximately five FP MS indels across the entire exome, in agreement with the FP cutoff used in MuTect¹⁶. To achieve this, we chose values of AIC $T_r = 8$ 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, because these are the most prevalent MS indel events in the genome and are also the most challenging to detect.

We first created virtual tumor data sets by using the same three WES data sets 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 by 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 with fractions $(1 - fr, fr)$.

We then used MSMuTect to detect somatic mutations by comparing the simulated tumor and the third copy of NA12878 (serving as the matched normal sample). We evaluated the sensitivity of MSMuTect to identify MS indels for various allele fractions and repeat lengths (**Supplementary Fig. 5**).

We evaluated MSMuTect by 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 $l = k \pm 1$. The sensitivity was highest for shorter MS loci (e.g., the sensitivity decreased from 98% for AAAAA (denoted A_5), to 75% for A_{12}) (**Supplementary Fig. 5**). Simulated MS indels with an allele frequency below 20% exhibited high rates of false negatives, probably because the allele fraction of ‘artificial’ MS indels generated by PCR exceeded the simulated MS indel fraction.

RNA validation. For the list of the seven significant MS loci, we manually compared the 161 MS indels found in the STAD cohort and the corresponding tumor RNA-seq data obtained from TCGA. An indel was confirmed if at least two RNA-seq reads supported the mutant MS allele (**Supplementary Table 3**).

MSI and POLE classification. For each sample, a score associated with *POLE* mutations and a score associated with MSI mutations were calculated on the basis of the ratio of signal mutations (i.e., mutations uniquely associated with the mutational process) to background mutations (other mutations). For *POLE*, the signal mutation⁶³ is C>A in the context TCT>TAT, 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 methyl-CpG dinucleotides (sometimes called the ‘aging’ signature) and the APOBEC-associated signatures⁶³. 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), according to previous analyses²¹, 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, which combined genes from the Cancer Gene Census⁶⁴ list with gene lists from other accepted sources and recent publications.

Diversity in normal samples. In MSMuTect, we identify the MS alleles in the normal samples before comparing them with tumor alleles. For each MS locus, we analyze the alleles across all normal samples and calculate its diversity, i.e., the fraction of normal samples with an allele different from that in the reference genome. For the significance analysis (for both MSMuSig and the search for new events in known cancer genes), we exclude loci with >10% diversity. Although this rationale is similar to the rationale for using a panel-of-normals comparison to exclude sites with either missed germline events or sequencing artifacts¹⁶, in MS loci, this approach may also identify sites that are more prone to MS indels and have a naturally higher mutation rate.

MSMuSig. MSMuSig 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 (**Fig. 3b,c**), whereas other covariates that are known to influence SNV rates (such as replication timing) have minimal effects on MS mutation rates. Thus, we separately estimated the background mutation frequency for each motif and repeat length in every tumor type.

We estimated the rates (and tested the significance) of loci containing at least one MS mutation across the analyzed cohort. We calculated these conditional rates (i.e., conditional on observing at least one event), because we observed broad variability in mutation rates with a significant enrichment of sites with no mutation. Estimating the mutation rate including these ‘stable’ sites would underestimate the overall background rate and hence expand 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, a value approximately six times higher than we would have expected (35 loci, $P < 10^{-16}$, one-tailed 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 exclusion of these ‘stable’ sites, there was still a high variability in mutation rates among MS loci with the same motif and repeat length, beyond

the variability that would be expected from a binomial distribution, assuming that 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 contain functionally relevant MS loci than exons, such as untranslated regions and introns (**Supplementary Fig. 9**).

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 quantile–quantile plots for the noncoding MS loci and coding MS loci are shown in **Supplementary Figures 10–12** (for different tumor types). There was no inflation of significantly mutated sites, and most MS loci followed the expected uniform P -value distribution (i.e., they resided close to the diagonal in the quantile–quantile plot).

Expression data. The RNA-seq-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 in which they were not available, we used \log_2 RPKM values.

Code availability. Code for MSMuTest and MSMutSig are supplied in **Supplementary Software 1 and 2**, respectively.

Data availability. The three whole-exome sequencing replicates of NA12878 and a full list of the MS indels including noncoding and germline heterozygous sites are available via a TCGA-protected workspace in FireCloud (<http://www.firecloud.org/>) upon request to the authors.

A **Life Sciences Reporting Summary** for this paper is available.

57. The Cancer Genome Atlas Data Portal. <https://tcga-data.nci.nih.gov/docs/publications/tcga/> (accessed 10 October, 2016).
58. 1000 Genomes Project Consortium. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
59. Mayer, C., Leese, F. & Tollrian, R. Genome-wide analysis of tandem repeats in *Daphnia pulex*: a comparative approach. *BMC Genomics* **11**, 277 (2010).
60. Ramos, A.H. *et al.* Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
61. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
63. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
64. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Experimental design

Sample size

Describe how sample size was determined.

We used all relevant samples from the TCGA.

Data exclusions

Describe any data exclusions.

We used only sequencing data from Illumina sequencers (the vast majority of cases).

Replication

Describe whether the experimental findings were reliably reproduced.

Not relevant

Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not relevant

Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not relevant

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

Confirmed

- ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ A statement indicating how many times each experiment was replicated
- ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- ☐ The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- ☐ Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

We used tools that were developed as part of this work and they are described in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Not relevant.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Not relevant.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Not relevant.

b. Describe the method of cell line authentication used.

Not relevant.

c. Report whether the cell lines were tested for mycoplasma contamination.

Not relevant.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

Not relevant.

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Not relevant.

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Not relevant.