ANALYSIS

nature medicine

Classification and characterization of microsatellite instability across 18 cancer types

Ronald J Hause[1], Colin C Pritchard[2], Jay Shendure[1,3] & Stephen J Salipante[2]

Microsatellite instability (MSI), the spontaneous loss or gain of nucleotides from repetitive DNA tracts, is a diagnostic phenotype for gastrointestinal, endometrial, and colorectal tumors, yet the landscape of instability events across a wider variety of cancer types remains poorly understood. To explore MSI across malignancies, we examined 5,930 cancer exomes from 18 cancer types at more than 200,000 microsatellite loci and constructed a genomic classifier for MSI. We identified MSI-positive tumors in 14 of the 18 cancer types. We also identified loci that were more likely to be unstable in particular cancer types, resulting in specific instability signatures that involved cancer-associated genes, suggesting that instability patterns reflect selective pressures and can potentially identify novel cancer drivers. We also observed a correlation between survival outcomes and the overall burden of unstable microsatellites, suggesting that MSI may be a continuous, rather than discrete, phenotype that is informative across cancer types. These analyses offer insight into conserved and cancer-specific properties of MSI and reveal opportunities for improved methods of clinical MSI diagnosis and cancer gene discovery.

MSI is a molecular tumor phenotype resulting from genomic hypermutability. The gain or loss of nucleotides from microsatellite tracts—DNA elements composed of short repeating motifs—is the diagnostic hallmark of MSI[1] and manifests as novel alleles of varying length[2]. These changes can arise from impairments in the mismatch repair (MMR) system, which limits correction of spontaneous mutations in repetitive DNA sequences[3,4]. MSI-affected tumors may, accordingly, result from mutational inactivation or epigenetic silencing of genes in the MMR pathway[2,3]. MSI is classically associated with colorectal cancers, for which it holds well-defined clinical implications[3]. However, MSI has been reported in diverse cancer types including endometrial, ovarian, gastric, and prostate cancer and glioblastoma[3,5,6]. Recent work suggests that MSI may be an actionable marker for immune-checkpoint-blockade therapy; clinical trials have demonstrated improved outcomes for patients with MSI-positive tumors treated with inhibitors of programmed cell death 1 (PD-1), presumably as a result of T lymphocyte recognition of neoantigens produced by somatic mutations[7,8]. However, mutations resulting from MSI can also drive oncogenesis, by inactivating tumor suppressor genes, for example[9]. These observations underscore the need for a more complete understanding of MSI.

MSI signatures may differ among cancer types; disparate loci may be preferentially unstable[5,10–14], MSI positivity may carry different prognostic values[11], and MSI may occur at different frequencies[5] across malignancies. However, these observations come from examination of dozens of loci in cohorts no larger than 100 individuals. Beyond limited studies restricted to four cancer types with established MSI phenotypes[10,15,16], variation in MSI among malignancies has not yet been evaluated systematically or on a genomic scale.

Molecular diagnosis of MSI is currently achieved by examining PCR products from a few (typically 5–7) informative microsatellite markers (MSI–PCR)[1,17]. Recently, our group and others[10,18–22] developed methods to infer MSI using massively parallel DNA-sequencing technologies, enabling interrogation of MSI with a breadth and quantitative precision not previously achievable. Here, we describe a robust approach for predicting MSI status independently of cancer type and use tumor exomes from the Cancer Genome Atlas (TCGA) Research Network to more comprehensively examine MSI across tumor types.

RESULTS

MSI classifier

From a total of 19,075,236 microsatellites computationally identified across the human genome, we included a subset of 516,876 loci (2.7%) that were within or adjacent to the exome capture baits used by TCGA, representing 95.9% of all coding microsatellites and 98.4% of microsatellites occupying splice sites (**Fig. 1a**,**b** and **Supplementary Table 1**). These loci were primarily mononucleotide repeats (**Supplementary Table 2**) and, as expected from our study design, fell disproportionately into intronic and coding regions compared to distributions observed genome-wide (**Fig. 1b** and **Supplementary Table 3**). Insufficient sequencing read depth precluded interrogation of all microsatellites for every specimen: 223,082 loci (43%) had sufficient coverage (≥30 reads) in both tumor and normal tissue for instability status to be inferred in at least half of the 5,930 total specimens.

For each locus we catalogued microsatellite allele lengths in tumor and patient-matched normal exomes (**Fig. 1c**, **Supplementary Fig. 1**, and **Supplementary Tables 4** and **5**) to identify and quantify MSI events. Using these instability calls, we designed a classifier to

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA. [3]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to S.J.S. (stevesal@uw.edu) or J.S. (shendure@uw.edu).

Received 3 May; accepted 29 August; published online 3 October 2016; corrected after print 19 July 2017; doi:10.1038/nm.4191

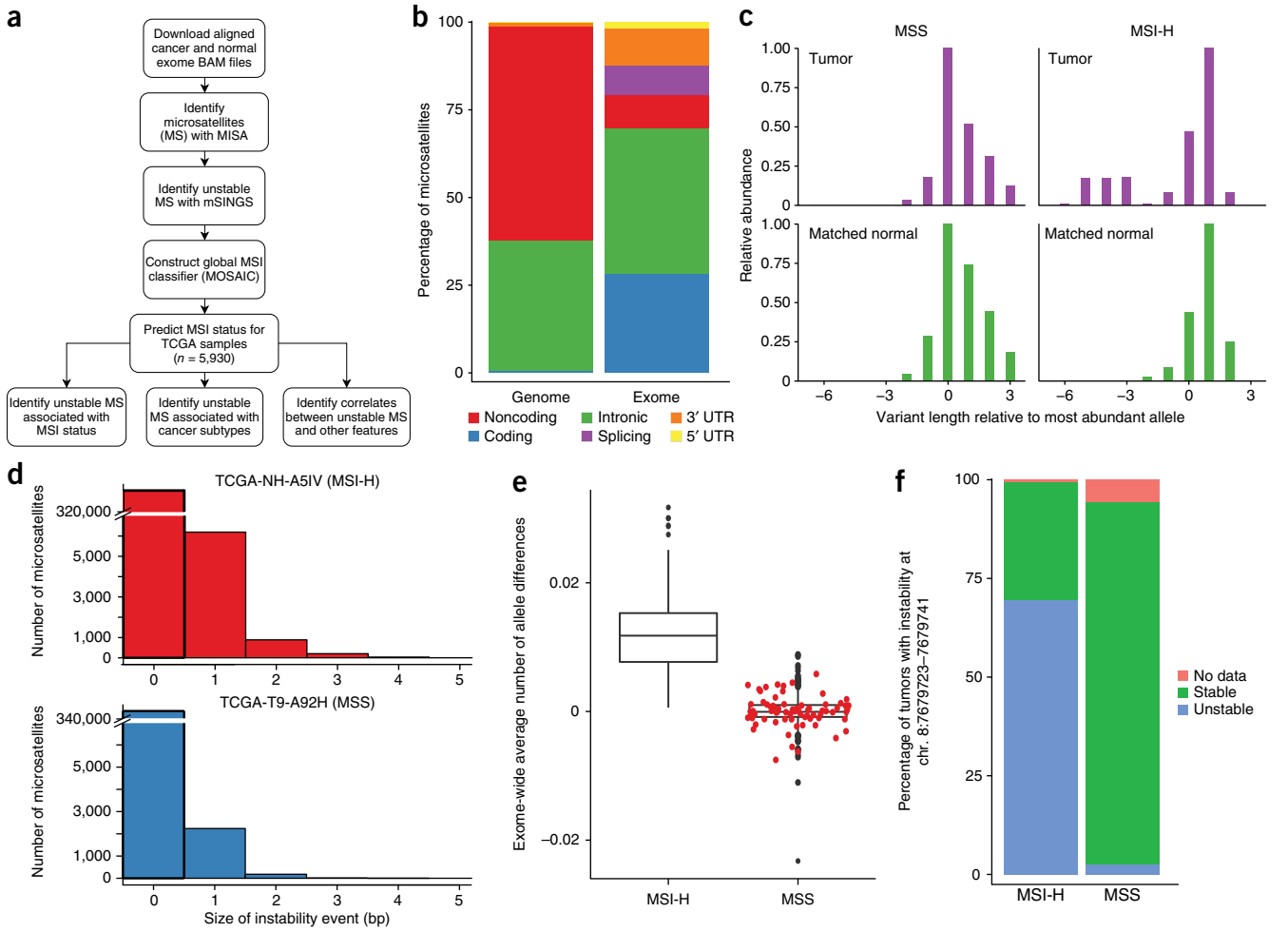

Figure 1 Evaluating MSI using exome-sequencing data. (**a**) Schematic of the approach used for analyzing MSI across TCGA exomes. MISA, microsatellite identification tool, mSINGS, microsatellite instability by next-generation sequencing. (**b**) Relative proportions of microsatellite loci within indicated genomic annotations across the whole genome and regions targeted by exome capture. Data represent computational identification and annotation of all microsatellites in the human reference genome and the subset within or immediately adjacent to TCGA exome capture baits. (**c**) Detection of MSI events from sequencing data. Representative virtual electropherograms[21] of a compound repeat at chr. 1:33145935–33145982 are illustrated for MSS and MSI-H cases, comparing the length and relative abundance of microsatellite alleles between tumors and patient-matched normal material. (**d**) Size of MSI events in representative MSI-H and MSS colon cancers. TCGA patient identifiers are indicated. (**e**) Correlation between MSI status (diagnosed using conventional clinical methods; MSI-H $n = 171$, MSS $n = 446$) and differences in global measurements of locus instability in tumor and paired normal specimens. Box boundaries indicate the interquartile range; center lines, medians; whiskers, values within 1.5 interquartile ranges of median; circles, extreme outliers. Red points represent MSI-L cancers ($n = 73$). (**f**) Proportion of MSI-H and MSS tumors with instability in a microsatellite locus located at chr. 8:7679723–7679741, within *DEFB105A/B*. This locus was the most significantly unstable microsatellite in MSI-H ($n = 171$) relative to MSS tumors ($n = 446$, $P = 2 \times 10^{-61}$, Fisher's exact test).

distinguish MSI-positive (MSI-high (MSI-H)) from MSI-negative (MSI-stable (MSS)) specimens independently of cancer type. Of all covariates tested across a cohort of colon, rectal, endometrial, and gastric tumors with available MSI–PCR results, the average total gain in the number of microsatellite alleles observed in a tumor relative to normal tissue across all microsatellite loci was the most significant feature separating MSI-H from MSS cancers (**Fig. 1d**,**e**; MSI-H median = 0.012, MSS median = -5.4×10^{-5}, $P = 9.4 \times 10^{-80}$, two-sided Wilcoxon rank-sum test). Related metrics, including the overall numbers of unstable microsatellites and variances in the allele number gain between tumor and normal, were also significantly different between MSI status groups (**Supplementary Fig. 2**; $P < 10^{-72}$, two-sided Wilcoxon rank-sum test). We also tested all microsatellite loci for discriminatory power to differentiate MSI-H from MSS samples and identified a locus within *DEFB105A* or *DEFB105B* (*DEFB105A/B*), chr. 8:7679723–7679741, as the most significantly unstable microsatellite in MSI-H tumors, as compared to MSS tumors (**Fig. 1f**; unstable in 119 of 171 MSI-H and 11 of 446 of MSS tumors, $P = 2 \times 10^{-61}$, two-sided Fisher's exact test). On the basis of these data, we created a parsimonious, weighted-tree microsatellite instability classifier (MOSAIC) for predicting MSI status using the most informative and independent features for classifying MSI—average gain of novel microsatellite alleles detected in a tumor specimen and, secondarily, locus instability within *DEFB105A/B* (**Supplementary Fig. 3a**). Incorporating additional covariates did not substantially improve the classifier (**Supplementary Fig. 3b**), nor did more sophisticated machine learning approaches. Compared with MSI–PCR, MOSAIC classified MSI-H from MSS cancers with 96.6% leave-one-sample-out cross-validation accuracy (95.8% sensitivity, 97.6% specificity) in a set of 617 specimens (128 MSS and 44 MSI-H for

colon; 63 MSS and 3 MSI-H for rectal; 169 MSS and 92 MSI-H for endometrial; 86 MSS and 32 MSI-H for stomach cancers). MOSAIC was discordant with clinical testing in classifying 11 of 171 MSI-H tumors (1 rectal and 10 endometrial) as MSS and 7 (1 rectal, 1 colon, and 5 endometrial) of 446 MSS cancers as MSI-H (**Supplementary Table 6**). Discordant classifications were primarily in endometrial cancers, which showed the smallest differences between MSI-H and MSS groups for all instability metrics measured. However, evidence suggests that many of these specimens were improperly classified by MSI–PCR: a review of accessory genetic and epigenetic data for somatic disruption of MSI-causative genes revealed that 7 of the 16 cases with complete metadata available were compatible with MOSAIC classifications but not MSI–PCR results (**Supplementary Table 6**). Furthermore, in terms of average number of gained microsatellite alleles and global burden of unstable microsatellites, discordant specimens were more consistent with MOSAIC classifications than with MSI–PCR testing (**Supplementary Fig. 4**).

Last, we evaluated whether differences in sequencing read depth between matched tumor and normal exomes or across microsatellite loci could confound our analysis. We observed no meaningful correlation between instability calls and these read depth metrics ($R^2 = 0.01$ and $\rho = -0.04$, respectively; **Supplementary Fig. 5**). Overall, these results demonstrate that we can accurately classify MSI status from tumor and matched-normal tissue exome-sequencing data.

Investigation of MSI-low phenotype

MSI-low (MSI-L) is a subcategory of MSI marked by instability at a minimal fraction of typed microsatellite markers. It is debated

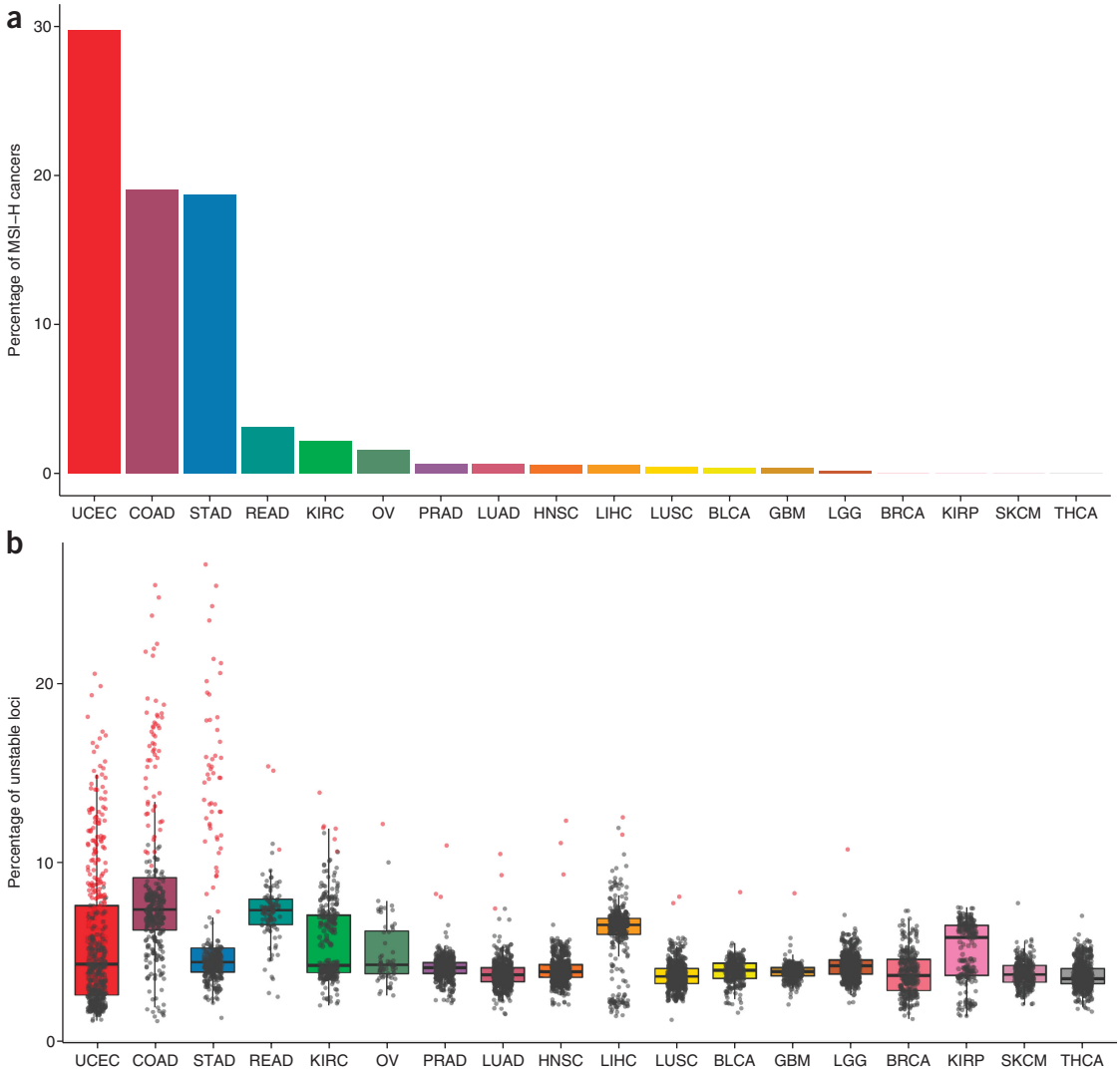

Figure 2 The landscape of MSI across TCGA exomes. (**a**) Inferred proportion of MSI-H tumors identified for each cancer cohort. (**b**) Distributions of the overall percentages of unstable microsatellite loci identified for each cancer type. Box boundaries indicate the interquartile range; center lines, medians; whiskers, values within 1.5 interquartile ranges of median. Overlaid points represent the number of unstable loci detected in individual tumor specimens; data for tumors classified as MSI-H are shown in red. UCEC, uterine corpus endometrial carcinoma ($n = 437$); COAD, colon adenocarcinoma ($n = 294$); STAD, stomach adenocarcinoma ($n = 278$); READ, rectal adenocarcinoma ($n = 96$); KIRC, kidney renal clear cell carcinoma ($n = 279$); OV, ovarian serous cystadenocarcinoma ($n = 63$); PRAD, prostate adenocarcinoma ($n = 463$); LUAD, lung adenocarcinoma ($n = 480$); HNSC, head and neck squamous cell carcinoma ($n = 506$); LIHC, liver hepatocellular carcinoma ($n = 338$); LUSC, lung squamous cell carcinoma ($n = 443$); BLCA, bladder urothelial carcinoma ($n = 253$); GBM, glioblastoma multiforme ($n = 262$); LGG, brain lower grade glioma ($n = 513$); BRCA, breast invasive carcinoma ($n = 266$); KIRP, kidney renal papillary cell carcinoma ($n = 207$); SKCM, skin cutaneous melanoma ($n = 268$); THCA, thyroid carcinoma ($n = 484$).

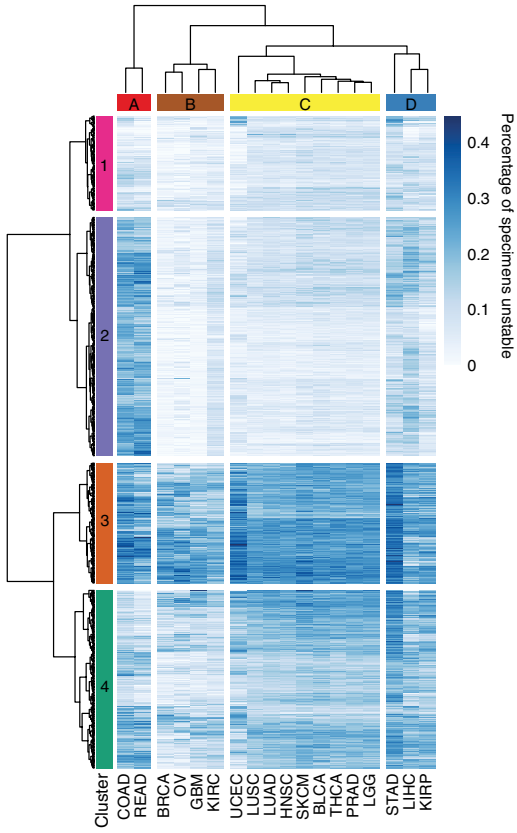

Figure 3 Cancer-specific signatures of MSI. Heatmap indicating the proportions of specimens within cancer types (columns) that were unstable at individual loci microsatellites (rows). Loci significant for differences among cancer types at FDR < 0.05 are shown. Colored microsatellite clusters (1–4, at left) denote groups of loci with similar instability trends based on Bayesian information criterion of the most likely model and number of clusters. Cancer types were also organized by hierarchical clustering into groups with similar patterns of MSI (A–D, top). UCEC, uterine corpus endometrial carcinoma; COAD, colon adenocarcinoma; STAD, stomach adenocarcinoma; READ, rectal adenocarcinoma; KIRC, kidney renal clear cell carcinoma; OV, ovarian serous cystadenocarcinoma; PRAD, prostate adenocarcinoma; LUAD, lung adenocarcinoma; HNSC, head and neck squamous cell carcinoma; LIHC, liver hepatocellular carcinoma; LUSC, lung squamous cell carcinoma; BLCA, bladder urothelial carcinoma; GBM, glioblastoma multiforme; LGG, brain lower grade glioma; BRCA, breast invasive carcinoma; KIRP, kidney renal papillary cell carcinoma; SKCM, skin cutaneous melanoma; THCA, thyroid carcinoma.

whether MSI-L is a distinct disease entity or an artifact of examining small numbers of loci through conventional MSI testing[23]. We therefore examined exome instability covariates in colon, rectal, endometrial, and stomach cancers clinically categorized as MSI-L. We observed no significant differences between MSI-L and MSS cancers in numbers of gained microsatellite alleles in tumor relative to normal tissue (P = 0.73, two-sided Wilcoxon rank-sum test), overall variation in allele number differences across all loci (P = 0.10), or total number of unstable microsatellites (P = 0.20; **Fig. 1f** and **Supplementary Fig. 2**). The lack of observable differences between these categories supports previous observations[10] and indicates that MSI-L tumors are consistent with MSS tumors in overall MSI burden. We reclassified MSI-L tumors as MSS for all subsequent analyses.

MSI status and landscape across different cancers

We broadly applied MOSAIC to assign MSI status for 5,930 tumor exomes from 18 cancer types (**Fig. 2a** and **Supplementary Tables 4** and **5**), enabling us to extend the analysis to 15 additional cancer types for which MSI status is not tested in clinical practice and to identify an additional 93 MSI-H samples. Cancer exomes contained a wide range of unstable microsatellites, from 87 to 9,032 (**Supplementary Table 5**). The average number of unstable sites varied considerably by cancer type, from a minimum of 765, for thyroid carcinomas, to a maximum of 2,315, for colon cancers. Similarly, the fraction of inferred MSI-H tumors also varied. The highest proportion of MSI-H cases occurred in cancer types that classically demonstrate MSI: endometrial (30%), colon (19%), and gastric (19%). Rectal cancers had a lower prevalence of MSI-H specimens (3%). Still lower, but detectable, frequencies of MSI-H were observed in 12 other cancer types; collectively, one or more individual MSI-H tumors were identified in 16 of the 18 cancer types examined. For several cancer types, including kidney papillary, kidney clear cell, and liver hepatocellular carcinomas, we observed a bimodal distribution in the proportion of unstable microsatellites for cancers classified as MSS (**Fig. 2b**), indicating trends in instability rates within MSI classifications.

As anticipated, we observed a strong correlation between predicted MSI status and the occurrence of somatic mutations or epigenetic silencing in MMR-pathway and DNA proofreading genes (odds ratio (OR) = 13.7 for having a somatic mutation in MSI-H malignancies compared with MSS, $P = 6 \times 10^{-64}$; **Supplementary Table 7**). Notably, these somatic alterations did not predict MSI-H status with high accuracy, suggesting contributions of additional factors to MSI. Despite the well-established role of mismatch repair gene *MLH1* silencing in MSI-H tumors[1], 8 of 98 tumors with *MLH1* silencing were classified as MSS by both MOSAIC and MSI–PCR.

To provide a more comprehensive view of the MSI landscape within and across cancer types, we next examined global patterns of microsatellite mutation using instability calls for individual loci. We included all specimens, irrespective of inferred MSI status, and restricted analysis to 92,385 microsatellites that were called in at least half of the samples across each of the 18 cancer types (**Supplementary Table 8**). No instability was observed at 57.4% of loci in any tumor. Of the sites that were unstable in at least 5% of specimens, hierarchical clustering distinguished four major groups (A–D) of cancer types having similar signatures of MSI (**Fig. 3**). Cancers that are canonically affected by MSI were distributed between three different groups: colon and rectal cancers exclusively comprised group A, whereas stomach cancers were placed in a separate category with liver hepatocellular and kidney renal carcinoma (D), and endometrial tumors were separately grouped with multiple other cancer types (C). Other malignancies, representing those with lower or no inferred incidences of MSI-H, were distributed among three groups (B, C, and D) but were disproportionately allocated to group C. All cancer types, including those entirely comprising MSS tumors, showed high frequencies of instability events at particular loci or groups of similarly mutated loci. The microsatellite loci were also partitioned by hierarchical clustering into four major divisions (1–4) that showed similar rates of instability across cancer groups (**Fig. 3**). We examined enrichment of gene ontologies and KEGG pathway annotations of factors harboring unstable microsatellites in each division and noted differences (**Supplementary Fig. 6** and **Supplementary Table 9**) but observed no obvious patterns of biological function.

Differences between MSS and MSI-H cancers

Because MSS tumors have a low baseline level of MSI[24], we examined whether MSS tumors mutate at the same loci as tissue-matched MSI-H tumors by comparing their relative frequencies of instability events at each microsatellite. For sufficient numbers, we focused on the four cancer types with the highest incidence of MSI-H samples (colon, rectal, endometrial, and stomach). Although both the frequency of instability events and the number of alternative microsatellite alleles were significantly elevated in MSI-H tumors, they tended to occur at the same loci that were unstable at lower frequencies in MSS cases (**Fig. 4a**,**b**); we observed a correlation between the frequency of MSI events in MSI-H and MSS malignancies of the same

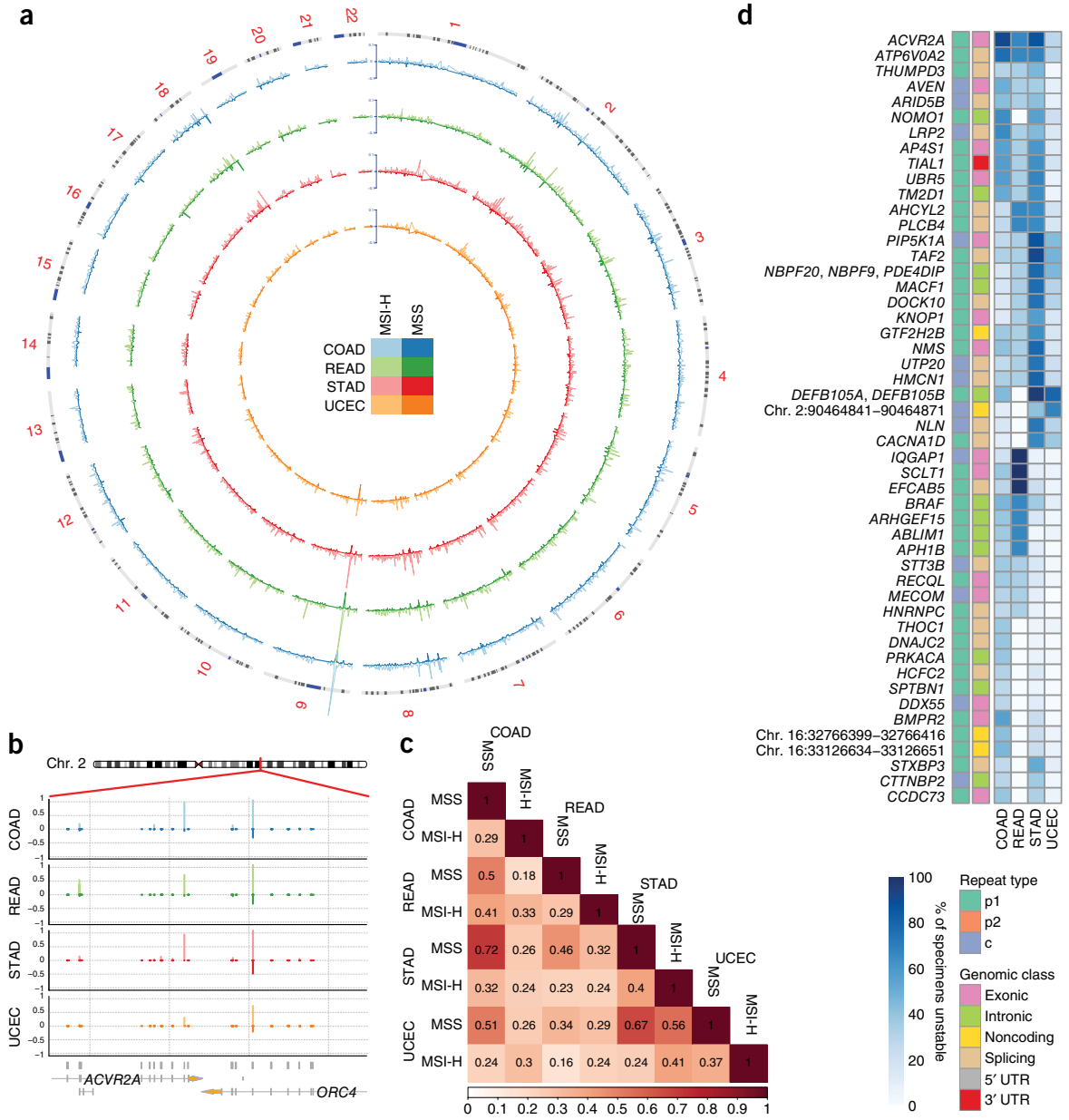

Figure 4 Signatures of MSI in MSI-H tumors. (**a**) Circos plot representing differences in the proportion of MSS and MSI-H tumors with instability at microsatellite loci across the genome. External ring denotes chromosomal position; internal rings indicate the average proportion of cancers with microsatellite locus instability for markers integrated across 1-Mb windows. For each cancer type, the proportion of tumors with instability within each window is indicated by the height of the trace, with MSI-H and MSS tumors for each cancer type plotted separately on different sides of the axis (positive and negative, respectively). The large peak on chromosome 9 reflects a single microsatellite in an intergenic region between *LINGO1* and *LOC401497* that was frequently unstable in both MSI-H and MSS cancers. (**b**) Representative region of chromosome 2 demonstrating differences in MSI profiles between MSI-H and MSS tumors and differences among MSI-H cancer types. The proportion of tumors with locus instability at each locus is indicated as in **a**. (**c**) Cosine similarity matrix comparing the sets of microsatellites unstable in at least half of tumors in each cancer subtype (stratified by MSI status and cancer type). (**d**) Heatmap indicating the proportion of specimens within MSI-H cancer types affected (columns) at the top 50 most differentially unstable microsatellite loci (rows). All loci were significant at FDR < 10^{-5}. For each microsatellite, repeat type and genomic annotation are indicated. p1, mononucleotide repeat; p2, dinucleotide repeat; c, compound repeat. UCEC, uterine corpus endometrial carcinoma (n = 437); COAD, colon adenocarcinoma (n = 294); STAD, stomach adenocarcinoma (n = 278); READ, rectal adenocarcinoma (n = 96).

Table 1 Ten most significant loci associated with MSI-H cancers

Locus coordinates	Proportion unstable (MSI-H)	Proportion unstable (MSS)	*P* value	*Q* value	OR	Genomic class	Gene(s)	Repeat sequence
Chr. 8:7679723–7679741	190/263 (72%)	173/5626 (3%)	9.19×10^{-191}	1.88×10^{-185}	81.76	Intronic	*DEFB105A*, *DEFB105B*	(A)9
Chr. 2:148683681–148683698	134/253 (52%)	30/5504 (<1%)	1.97×10^{-168}	2.02×10^{-163}	203.24	Coding	*ACVR2A*[a]	(A)8
Chr. 8:7346862–7346880	188/263 (71%)	274/5625 (4%)	3.55×10^{-161}	2.42×10^{-156}	48.849	Intronic	*DEFB105A*, *DEFB105B*	(T)9
Chr. 17:56435156–56435172	112/243 (46%)	10/5557 (<1%)	6.86×10^{-154}	3.51×10^{-149}	471.32	Coding	*RNF43*[a]	(C)7
Chr. 3:51417599–51417615	109/233 (46%)	24/4824 (<1%)	3.55×10^{-133}	1.46×10^{-128}	174.42	Coding	*DOCK3*[a]	(C)7
Chr. 7:74608736–74608758	230/257 (89%)	873/4882 (17%)	3.17×10^{-129}	1.08×10^{-124}	39.1	ncRNA Intronic	*GTF2IP1*, *LOC100093631*	(T)13
Chr. 11:120350632–120350654	104/264 (39%)	29/5579 (<1%)	1.31×10^{-121}	3.82×10^{-117}	124.15	Intronic	*ARHGEF12*[a]	(T)8(C)5
Chr. 16:14983087–14983105	100/264 (37%)	25/5643 (<1%)	5.99×10^{-119}	1.53×10^{-114}	136.12	Intronic	*NOMO1*[a]	(A)9
Chr. 1:151196698–151196722	127/264 (48%)	110/5643 (1%)	6.84×10^{-119}	1.56×10^{-114}	46.56	Coding	*PIP5K1A*[a]	(T)9(C)6
Chr. 1:200594037–200594054	98/250 (39%)	23/5249 (<1%)	2.89×10^{-117}	5.91×10^{-113}	145.57	Intergenic	*KIF14*[a] (dist. = 4,175 bp), *DDX59* (dist. = 19,111 bp)	(T)8

Dist., distance from microsatellite to indicated gene.
[a]Gene is implicated in oncogenesis (**Supplementary Table 12**).

cancer type and also across types (ρ = 0.28–0.53), indicating related instability patterns within and across malignancies. A representative example is provided by two loci in neighboring genes *ACVR2A* (chr. 2:148683681–148683698) and *ORC4* (chr. 2:148701095–148701119) (**Fig. 4b**). *ACVR2A* contains a coding mononucleotide microsatellite that is unstable in 28–90% of MSI-H samples (**Supplementary Table 10**), depending on the cancer type, but in only 0–6% of MSS cancers. *ORC4* harbors a mononucleotide repeat in a splicing region that is also unstable in 67–100% of MSI-H tumors but in only 19–44% of MSS samples.

To examine differences between MSS and MSI-H categories, we focused on microsatellites that were unstable in at least 25% of the samples within each cancer subtype and typable in all specimens. We computed cosine similarities between sets of all frequently unstable sites between each group (**Fig. 4c**). Colon, rectal, gastric, and endometrial MSI-H cancers intersected at a large fraction of their frequently unstable microsatellites, with tissue-matched MSS cancers sharing a smaller subset of those loci. MSS tumors from different tumor types showed substantially less overlap. Taken together, these findings indicate that MSI patterns in tissue-matched MSI-H and MSS cancers are related and follow consistent patterns, but MSI-H cancers share overall similarities in their most frequently unstable sites.

Differences among MSI-H cancers

To compare MSI among different MSI-H cancer types, we examined only cancer types with the highest MSI-H prevalence to lend sufficient power for statistically meaningful comparisons. As was observed for the entire collection of specimens (**Fig. 3**), separate MSI-H cancer types showed individualized signatures of instability at a subset of microsatellite loci (**Fig. 4d**). In total, 2,685 of the 3,296 microsatellites unstable in at least 5% of MSI-H cancers were differentially unstable in at least one cancer type at an FDR < 0.05 (**Supplementary Table 10**). These differentially unstable microsatellites included several in *NIPBL*, *TCF4*, and *PTEN*, among other genes reported as mutational targets of MSI[25,26]. An example is again provided by the microsatellites in *ACVR2A* and *ORC4* (**Fig. 4b**): the former was unstable in 90% of colon, 67% of rectal, and 87% of stomach MSI-H tumors, but only 28% of endometrial MSI-H tumors, and the latter was unstable in 97% of colon, 67% of endometrial, and 100% of rectal and stomach MSI-H tumors investigated.

To explore the functional consequences of different instability signatures among MSI-H cancer types, we examined factors that were uniquely unstable in one cancer type (**Supplementary Fig. 7**). Uniquely unstable factors in colon and rectal cancers shared overlap in multiple aggregated functional categories, although cancer-type-specific differences were observed among assorted cellular functions. Stomach adenocarcinomas were uniquely and highly enriched for instability in ion-binding genes while demonstrating instability in several categories frequently observed for MSI-H colon and rectal tumors. Endometrial cancers were exclusively enriched for uniquely unstable sites in protein complex binding genes, without overlap in categories identified for other cancer types, although the small number of endometrial-cancer-specific unstable sites limited our power for ascertaining such ontological enrichments.

Properties of unstable microsatellites

We investigated features associated with unstable loci by associating various intrinsic properties, annotations, and metrics with the likelihood of locus instability. After stratifying by repeat composition and MSI status, we found compound microsatellites to be more preferentially unstable than other repeat types, with 11.7% and 5.3% of those loci unstable in more than 20% of MSI-H and MSS samples, respectively (**Supplementary Fig. 8a**). Intrinsic length of the microsatellite tract had bearing on instability frequency, with a maximum occurring around 16 repeat units in length (**Supplementary Fig. 8b**). When loci were stratified by their genomic annotations (**Supplementary Fig. 8c**), microsatellites in coding regions were less likely to be unstable in at least one sample (OR = 0.87, $P = 2.3 \times 10^{-57}$). By contrast, microsatellites in splice sites were more likely to be unstable (OR = 1.37, $P = 2.2 \times 10^{-82}$). We compared primary sequence enrichments of microsatellites unstable in at least one cancer (**Supplementary Fig. 8d**) and observed no significant differences among MSI-H cancer types (**Supplementary Fig. 9a**). However, CA and GA dinucleotide repeats were the most likely to be unstable overall. We also observed variability in the likelihood of instability at CpG sites, which probably reflects their functional importance in gene regulation. We observed significant enrichments for instability at DNase hypersensitivity sites ($P = 0.01$), conserved transcription factor binding sites ($P = 3 \times 10^{-6}$), and evolutionarily conserved genomic regions ($P = 6 \times 10^{-9}$; **Supplementary Fig. 9b**). Last, we tested for a correlation between the average frequency of locus instability within 1-Mb windows across individuals and DNA replication timing[10,27] but found no significant associations (**Supplementary Fig. 9c**).

Unstable microsatellites in cancer-associated genes

To identify elements common to a generalizable MSI-H signature across cancer types, we tested for loci that were significantly more

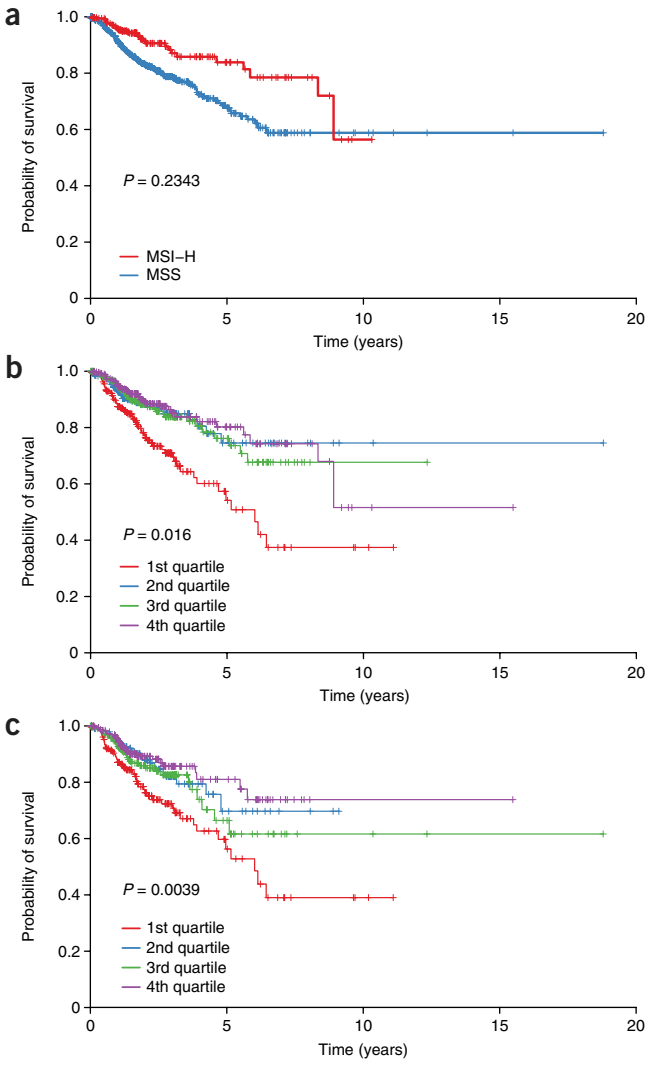

Figure 5 Global MSI load and patient survival. (**a**) Patient survival aggregated for endometrial, stomach, colon, and rectal cancers, stratified by inferred MSI status (MSS $n = 864$, MSI-H $n = 241$). (**b**) Patient survival for the same tumors in **a** as a function of the proportion of unstable microsatellites detected, grouped by quartile. (**c**) Patient survival for MSS cancers from **a**, grouped by quartile. *P* values in **b** and **c** represent the significance of the continuous variable of the proportion of unstable microsatellites per sample as determined from likelihood ratio tests. Significance in all panels was assessed after correcting for age, sex, radiation therapy status, and cancer type.

likely to be unstable in all MSI-H tumors ($n = 264$) than in all MSS cancers ($n = 5{,}666$). Of the 204,797 microsatellites with sufficient coverage to be called in at least half of MSI-H and MSS samples across cancer types, 17,564 sites within 6,882 unique genes were significant at an FDR < 0.05 (**Supplementary Fig. 8e** and **Supplementary Table 11**), indicating that a subset of markers are reliably unstable across cancer types and may represent common genomic lesions in MSI-H malignancies.

We noted that many recurrently unstable loci in MSI-H tumors (**Table 1**) involved cancer-associated genes, including coding regions in tumor suppressor genes *ACVR2A* and *RNF43*, which are frequent and validated targets of mutation in MSI-H cancers[28,29]. We explored a possible correlation of instability events with occurrence in genes participating in oncogenic pathways[9]. Using permutation testing, we tested whether recurrently unstable loci (**Supplementary Table 10**) were more likely to occur in genes registered in the COSMIC cancer gene census[30] and observed that microsatellites located in genes with known involvement in oncogenesis were significantly more likely to be unstable (**Supplementary Fig. 8f**; OR = 1.51, $P < 10^{-4}$). Moreover, a review of the literature for the genes harboring or proximal to the top 100 most significantly mutated loci in MSI-H cancers showed that 58 are in or near genes with previously established cancer-related biological functions (**Supplementary Table 12**). Furthermore, 25 of 27 known recurrent mutational targets of colorectal cancer MSI[26] examined in our study contained loci that were significantly unstable in MSI-H relative to MSS samples at an FDR < 0.05 ($P = 5 \times 10^{-9}$).

Patient survival and MSI burden

MSI-H status is associated with modestly improved patient survival in colorectal cancers[31]. We therefore examined whether there was a general correlation between MSI-H classification and survival outcome across cancer types, after correcting for covariates. We observed a weak association between MSI status and survival outcome when considering in aggregate the four cancer types with the highest incidence of MSI-H ($P = 0.23$, hazard ratio (HR) for MSI-H = 0.79; **Fig. 5a**). We next evaluated whether the global burden of unstable microsatellites would correlate with survival when treated as a continuous variable independently of MSI status and observed a stronger, more significant positive correlation with survival ($P = 0.02$, HR per increase of 100 unstable sites = 0.984; **Fig. 5b**). Given that MSI-H samples showed, on average, approximately 2,100 more unstable sites than MSS samples, this would equate to a HR of 0.72 for MSI-H. Furthermore, the association between the number of unstable sites and patient survival was more pronounced in MSS samples alone ($P = 0.004$, HR per increase of 100 unstable sites = 0.959; **Fig. 5c**). This observation led us to question whether the metric would also be prognostic of patient outcome in cancer types for which MSI is not typically evaluated. Although no significant effect was observed when cancer types were examined in aggregate, for individual cancer types we observed positive trends between prognosis and instability burden in uterine, endometrial, rectal, colon, stomach, and thyroid cancer and lower-grade glioma (**Supplementary Fig. 10**). Limited sample sizes for each cancer type restrict power for establishing the significance of these trends.

Last, we tested whether MSI was high in cancers that had progressed by quantifying instability events in primary and metastatic tumors within cancer types. We examined cancers for which multiple patient samples from metastatic disease were available, including seven patient-matched metastatic and primary breast tumors, seven patient-matched metastatic and primary thyroid tumors, and six primary and 174 metastatic melanoma cases from unrelated patients. All were MSS. The fractions of unstable loci were not significantly different between metastatic and primary tumors (median percentage unstable for each group = 0.37%, $P = 0.13$, nested ANOVA; **Supplementary Fig. 11**), which suggests that MSI is not associated with likelihood of metastasis, although additional samples will be necessary to substantiate this observation.

DISCUSSION

To explore the landscape of MSI in different cancers, we developed MOSAIC for ascertaining MSI status from tumor–normal tissue pairs examined with exome-sequencing data. Our approach leverages the observation that MSS tumors have a lower baseline level of instability

events than MSI-H tumors, which enables MSI classifications to be distinguished on the basis of global MSI calls. MOSAIC corrects for class imbalance in its cross-validation training procedure (an approximately 3:1 MSS-to-MSI-H ratio), allowing predictions in new cancer types to be made without prior assumption about the expected prevalence of MSI-H tumors. Although we noted a few discrepancies between our classifier and conventional MSI typing, genomic data suggest that these represent false positive and false negative outcomes from clinical typing[32,33] and that discordant results are more consistent with MOSAIC classifications.

Most cancer types examined (14 of 18) included one or more MSI-H representatives, suggesting that MSI may be a generalized cancer phenotype. The identification of infrequently occurring MSI-H tumors from cancer types conventionally associated with MSI confirms published reports[6,34–40]. Notably, most cancer types, even those for which there were few or no examples with the MSI-H phenotype in our cohort, showed a high frequency of MSI at restricted subsets of loci. This observation raises the possibility that findings[41] of MSI in some cancer types may reflect artifacts from typing local mutational hot spots by conventional methods rather than a global instability phenotype.

Microsatellite mutations occurring within the coding regions, introns, or untranslated regions of genes may positively or negatively influence gene expression or protein function by affecting changes in transcription or gene splicing[9,15,42–44]. We observed a depletion of unstable microsatellites in exons, transcription factor binding sites, and evolutionarily conserved genomic regions, consistent with purifying selection against mutations with biologically functional consequences[10]. Nevertheless, regulatory alterations for some targets may confer selective growth advantages to cancer cells, and unstable microsatellite loci have been speculated to fall within genes implicated in oncogenesis and to participate in the evolution of MSI-H cancers[9,13,14,16,45–47]. For unstable microsatellites observed in genic regions, our data support the idea that they preferentially accumulate in genes involved in carcinogenesis or tumor survival and therefore probably serve as drivers of cancer evolution. Differences in patterns of MSI among cancer types may consequently reflect different positive and negative selective pressures experienced during carcinogenesis. We observed that frequently unstable microsatellites in MSI-H malignancies are preferentially located in known cancer-associated genes, supporting this view and suggesting that there may be an underappreciated contribution of MSI in generating cancer-driving mutations. Moreover, roughly half of unstable microsatellites fall within genes not previously reported to be involved in cancer, including several intergenic loci, raising the possibility that these microsatellites also function as cancer drivers. Although functionally evaluating newly implicated factors is outside the scope of this work, many of the differences between MSS and MSI-H tumors are pronounced, and these data illustrate the utility of microsatellite analysis of exome-sequencing data as a primary approach for identifying cancer-relevant genes. Identification of features that are recurrently affected by MSI is complementary to methods that highlight genes on the basis of their recurrent somatic coding sequence mutations[48].

Although differences in selection during carcinogenesis may account for much of the variability in instability rates observed among microsatellite markers, we also observed significant correlations with more generalized properties of the loci themselves. We observed a weak but significant correlation between elevated MSI rates and loci occupying DNase-hypersensitivity sites, supporting earlier work[10] and indicating that instability events are enriched within euchromatic regions. Other factors, including repeat composition and locus length, affected instability[44]. It is likely that local nucleotide sequence or secondary structure surrounding repeats also define the inherent instability of a locus[44,49].

Consistent with other genomic studies[10,21], we found no evidence that tumors classified as MSI-L are a distinct disease group. This conclusion supports the view that MSI-L is a technical artifact reflecting a low background frequency of MSI in tumors with intact MMR systems[1,24]. Nevertheless, specimens in our study spanned a continuum of observed instability, and, at their extremes, tumors classified as MSI-H and MSS showed some overlap in their overall burden of unstable microsatellites. In general, we observed that the number of unstable microsatellite loci in a tumor exome correlated with patient survival when considered as a continuous metric better than conventional MSI-H or MSS classification alone. This result may reflect a link between MSI events and the production of cancer neoantigens that can be recognized as 'non-self' by the immune system[7,50]. Although the effect sizes we observed were smaller because of our limited cohort sizes, they are consistent with values reported in larger cohorts[31]. These findings suggest that, when sufficient numbers of loci are considered, the MSI phenotype may be a more continuous phenotype than previously appreciated—indeed, the global burden of MSI within MSS samples alone was prognostic of patient outcome. Because this continuous distribution of global instability is more indicative of patient survival independently of conventional MSI classification, it may prove more informative in the clinical management and treatment of cancer[7].

The existence of cancer-specific MSI landscapes and the potential predictive power of MSI as a continuous metric have implications for the molecular diagnosis of MSI in clinical practice: because current assays are optimized for the detection of MSI in colon and rectal cancers[17], they may not detect instability events effectively, or at all, in other cancer types. The behavior of any particular microsatellite locus can vary greatly across cancers, and loci that are inherently stable in one cancer type may be frequently mutated in another. Because MOSAIC for genome-scale MSI classification is more comprehensive and less prone to cancer-type-specific biases, it may serve as a better clinical strategy for pan-cancer MSI determination and ascertainment of instability burden.

Microsatellites are preferentially located in noncoding regions of the genome, and we anticipate that the future availability of more cancer whole-genome sequences will provide an improved understanding of the overall genomic landscape of MSI in different malignancies. As suggested by our study, such data may implicate novel, noncoding oncogenic motifs that affect gene regulation and will yield further insights into potentially important genomic sites involved in carcinogenesis.

METHODS

Methods and any associated 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 A. McKenna and other members of S.J.S. and J.S.'s laboratories for helpful advice and assistance. This work was supported in part by the Damon Runyon Cancer Research Foundation (DRG-2224-15 to R.J.H.), a Congressionally Directed Medical Research Program (PC131820 to C.C.P.), and a Young Investigator Award from the Prostate Cancer Foundation (to C.C.P.).

AUTHOR CONTRIBUTIONS
R.J.H. and S.J.S. conceived the work and designed and performed the analyses; R.J.H., S.J.S., C.C.P., and J.S. interpreted results; and R.J.H. and S.J.S. wrote the paper with input from C.C.P. and J.S.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. de la Chapelle, A. & Hampel, H. Clinical relevance of microsatellite instability in colorectal cancer. *J. Clin. Oncol.* **28**, 3380–3387 (2010).
2. Murphy, K.M. *et al.* Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. *J. Mol. Diagn.* **8**, 305–311 (2006).
3. Vilar, E. & Gruber, S.B. Microsatellite instability in colorectal cancer—the stable evidence. *Nat. Rev. Clin. Oncol.* **7**, 153–162 (2010).
4. Oki, E., Oda, S., Maehara, Y. & Sugimachi, K. Mutated gene-specific phenotypes of dinucleotide repeat instability in human colorectal carcinoma cell lines deficient in DNA mismatch repair. *Oncogene* **18**, 2143–2147 (1999).
5. Boland, C.R. *et al.* A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res.* **58**, 5248–5257 (1998).
6. Pritchard, C.C. *et al.* Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. *Nat. Commun.* **5**, 4988 (2014).
7. Le, D.T. *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
8. Timmermann, B. *et al.* Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole-exome next-generation sequencing and bioinformatics analysis. *PLoS One* **5**, e15661 (2010).
9. Woerner, S.M. *et al.* SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. *Nucleic Acids Res.* **38**, D682–D689 (2010).
10. 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).
11. Onda, M. *et al.* Microsatellite instability in thyroid cancer: hot spots, clinicopathological implications, and prognostic significance. *Clin. Cancer Res.* **7**, 3444–3449 (2001).
12. Forgacs, E. *et al.* Searching for microsatellite mutations in coding regions in lung, breast, ovarian and colorectal cancers. *Oncogene* **20**, 1005–1009 (2001).
13. Duval, A. *et al.* Target gene mutation profile differs between gastrointestinal and endometrial tumors with mismatch repair deficiency. *Cancer Res.* **62**, 1609–1612 (2002).
14. Mori, Y. *et al.* Instabilotyping reveals unique mutational spectra in microsatellite-unstable gastric cancers. *Cancer Res.* **62**, 3641–3645 (2002).
15. Sonay, T.B., Koletou, M. & Wagner, A. A survey of tandem repeat instabilities and associated gene expression changes in 35 colorectal cancers. *BMC Genomics* **16**, 702 (2015).
16. Yoon, K. *et al.* Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome Res.* **23**, 1109–1117 (2013).
17. Bacher, J.W. *et al.* Development of a fluorescent multiplex assay for detection of MSI-high tumors. *Dis. Markers* **20**, 237–250 (2004).
18. Lu, Y., Soong, T.D. & Elemento, O. A novel approach for characterizing microsatellite instability in cancer cells. *PLoS One* **8**, e63056 (2013).
19. McIver, L.J., Fonville, N.C., Karunasena, E. & Garner, H.R. Microsatellite genotyping reveals a signature in breast cancer exomes. *Breast Cancer Res. Treat.* **145**, 791–798 (2014).
20. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
21. Salipante, S.J., Scroggins, S.M., Hampel, H.L., Turner, E.H. & Pritchard, C.C. Microsatellite instability detection by next generation sequencing. *Clin. Chem.* **60**, 1192–1199 (2014).
22. Huang, M.N. *et al.* MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci. Rep.* **5**, 13321 (2015).
23. Pawlik, T.M., Raut, C.P. & Rodriguez-Bigas, M.A. Colorectal carcinogenesis: MSI-H versus MSI-L. *Dis. Markers* **20**, 199–206 (2004).
24. Laiho, P. *et al.* Low-level microsatellite instability in most colorectal carcinomas. *Cancer Res.* **62**, 1166–1170 (2002).
25. Kim, M.S., An, C.H., Chung, Y.J., Yoo, N.J. & Lee, S.H. NIPBL, a cohesion loading factor, is somatically mutated in gastric and colorectal cancers with high microsatellite instability. *Dig. Dis. Sci.* **58**, 3376–3378 (2013).
26. Boland, C.R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073–2087 (2010).
27. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
28. Jung, B. *et al.* Loss of activin receptor type 2 protein expression in microsatellite unstable colon cancers. *Gastroenterology* **126**, 654–659 (2004).
29. Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46**, 1264–1266 (2014).
30. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
31. Samowitz, W.S. *et al.* Microsatellite instability in sporadic colon cancer is associated with an improved prognosis at the population level. *Cancer Epidemiol. Biomark. Prev.* **10**, 917–923 (2001).
32. Hampel, H. *et al.* Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. *Cancer Res.* **66**, 7810–7817 (2006).
33. Goel, A., Nagasaka, T., Hamelin, R. & Boland, C.R. An optimized pentaplex PCR for detecting DNA mismatch repair–deficient colorectal cancers. *PLoS One* **5**, e9393 (2010).
34. Altavilla, G., Fassan, M., Busatto, G., Orsolan, M. & Giacomelli, L. Microsatellite instability and hMLH1 and hMSH2 expression in renal tumors. *Oncol. Rep.* **24**, 927–932 (2010).
35. Martinez, R. *et al.* Low-level microsatellite instability phenotype in sporadic glioblastoma multiforme. *J. Cancer Res. Clin. Oncol.* **131**, 87–93 (2005).
36. Jensen, K.C. *et al.* Microsatellite instability and mismatch repair protein defects in ovarian epithelial neoplasms in patients 50 years of age and younger. *Am. J. Surg. Pathol.* **32**, 1029–1037 (2008).
37. Kazachkov, Y. *et al.* Microsatellite instability in human hepatocellular carcinoma: relationship to p53 abnormalities. *Liver* **18**, 156–161 (1998).
38. Dacic, S., Lomago, D., Hunt, J.L., Sepulveda, A. & Yousem, S.A. Microsatellite instability is uncommon in lymphoepithelioma-like carcinoma of the lung. *Am. J. Clin. Pathol.* **127**, 282–286 (2007).
39. Field, J.K. *et al.* Microsatellite instability in squamous cell carcinoma of the head and neck. *Br. J. Cancer* **71**, 1065–1069 (1995).
40. Eckert, A. *et al.* Microsatellite instability in pediatric and adult high-grade gliomas. *Brain Pathol.* **17**, 146–150 (2007).
41. Amira, N. *et al.* Microsatellite instability in urothelial carcinoma of the upper urinary tract. *J. Urol.* **170**, 1151–1154 (2003).
42. Dorard, C. *et al.* Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. *Nat. Med.* **17**, 1283–1289 (2011).
43. Shin, N. *et al.* Identification of frequently mutated genes with relevance to nonsense-mediated mRNA decay in the high microsatellite instability cancers. *Int. J. Cancer* **128**, 2872–2880 (2011).
44. Shah, S.N., Hile, S.E. & Eckert, K.A. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res.* **70**, 431–435 (2010).
45. Woerner, S.M. *et al.* Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene* **22**, 2226–2235 (2003).
46. Duval, A. *et al.* Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. *Hum. Mol. Genet.* **10**, 513–518 (2001).
47. Imai, K. & Yamamoto, H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. *Carcinogenesis* **29**, 673–680 (2008).
48. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature* **505**, 495–501 (2014).
49. Eckert, K.A. & Hile, S.E. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.* **48**, 379–388 (2009).
50. Mlecnik, B. *et al.* Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability. *Immunity* **44**, 698–711 (2016).

ONLINE METHODS

Exome microsatellite data. Exome data for all specimens (tumors and patient-matched normal blood) were obtained from the TCGA Research Network (http://cancergenome.nih.gov/; **Supplementary Table 5**) as alignments against hg19. Researchers were not blinded to the MSI status of specimens where those data were available. We identified all autosomal microsatellite tracts with repeating subunits of 1–5 bp in length and comprising 5 repeats or more in the human reference genome (GRCh37/hg19) using MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html) and padded their start and stop coordinates by 5 bp. 10 bp or fewer were permitted between repeats for adjacent microsatellites to be combined into single loci, termed either 'complex' (c*) if comprised of microsatellites with different repeat subunit lengths or 'compound' (c) if comprised of disparate repeats with the same repeat length. Microsatellites directly tiled by the NimbleGen SeqCap_EZ_Exome_v3 capture design (which was used by TCGA) and those within 50 bp of a capture bait were retained. Repeat features were annotated using ANNOVAR[51] (24 February 2014 release).

Calling unstable microsatellite loci. Primary analysis of microsatellite loci was performed in each specimen to determine stability using mSINGS as previously described[21]. Briefly, we evaluated the number of sequence reads of different lengths present within each of the identified microsatellite markers, then expressed the relative abundance of individual lengths for a microsatellite as the fraction of reads supporting that length normalized to the number of reads counted for the most frequently occurring length at that locus. Microsatellite tract lengths at <5% relative abundance were discarded. Although identified length polymorphisms may include some reproducible artifacts resulting from slippage during PCR amplification, their total number is proportional to the actual number of microsatellite alleles present at a locus[21], and in comparative analysis of genetically related tumor–normal pairs such artifacts are well controlled. Instability at each locus was subsequently defined in two ways: (i) the high-sensitivity approach, in which identification was performed by comparing the absolute number of lengths identified between tumor and paired normal specimens, and the locus was considered unstable if one or more additional lengths for a microsatellite were detected from the tumor; and (ii) the high-specificity approach, in which Kolmogorov–Smirnov scores were calculated when comparing the normalized distribution of lengths for tumor and paired normal specimens, considering any difference less than $P = 0.05$ to signify locus instability[10]. We determined the latter method to be overly conservative (a median of only 5 unstable sites were called per MSI-H cancer), and therefore did not implement it in practice. Accordingly, the burden of unstable sites identified in our study was considerably higher than approximated in other work[10], probably because of the greater sensitivity-to-specificity tradeoff of our approach.

Constructing MOSAIC from sequencing-based locus instability calls. We examined data from colon, rectal, stomach, and endometrial cancer exomes (**Supplementary Tables 4** and **5**) for which clinical MSI status was available from standard diagnostic methods[17]. We observed that the average size of instability events (i.e., the length of alternate microsatellite alleles) was greater in MSI-H than MSS tumors (**Fig. 1d**,**e**; $P = 9 \times 10^{-80}$, two-sided Wilcoxon rank-sum test). Clinical MSI–PCR results (MSI-H, MSI-L, and MSS) were obtained from TCGA. The average gain in unique alleles in tumor relative to matched normal tissue across all interrogated microsatellites (peak_avg), variation in allele gain (peak_var), total number of unstable sites defined by the high-sensitivity method (num_unstable), and proportion of callable unstable sites (prop_unstable) were calculated for each sample. Furthermore, we tested for the power of each microsatellite locus to differentiate between MSI-H and MSS tumors using Fisher's exact tests and identified a locus within *DEFB105A/B*, chr. 8:7679723–7679741, as the most significantly unstable microsatellite in MSI-H relative to MSS tumors (defbsite). These features, along with the top 100 most significantly unstable microsatellites in MSI-H relative to MSS tumors, were then used to predict clinical MSI-H or MSS diagnosis by recursive partitioning classification trees or random forests implemented using the rpart v4.1-10, randomForest v4.6-12, and caret v6.0.62 packages in R v3.2.1. Leave-one-sample-out cross-validation was used to learn the optimal features and parameters for predicting MSI status, interrogating a grid search space of 0, 0.001, 0.01, 0.1, 0.45, and 0.95 complexity parameters (cp) with the minimum number of observations in any terminal node (minbucket) set to 6 and the maximum depth of any node of the final tree set to 3 for recursive partitioning, and 2 and 3 randomly sampled variables as candidates at each split (mtry) with 1,000 trees for random forests. Weights were included to correct for class imbalances in the training data (MSI-H $n = 171$, MSS $n = 446$), and the optimal parameters selected were cp = 0.001 and mtry = 2. Notably, peak_avg and debsite were selected by recursive feature selection using decision trees as the most significant two features for inclusion in the final model; incorporating more than two covariates did not significantly improve the classifier (**Supplementary Fig. 3b**). The final models achieved 96.6% (rpart) and 96.4% (randomForest) accuracy. The more accurate and parsimonious rpart model was used to predict MSI status across all remaining cancer samples.

Identifying uniquely unstable microsatellites in MSI-H cancers. For each of the 204,797 microsatellite loci called in at least half of MSI-H and MSS cancers ($n > 132$ and $n > 2{,}833$, respectively), we performed two-sided Fisher exact tests comparing the ratios of individuals for which the site was unstable in MSI-H samples to the ratio of individuals for which the site was unstable in MSS samples. FDR values were estimated using Storey's q-value method, with a q-value < 0.05 considered significant.

Determining cancer-specific microsatellite sites. Multiple proportions tests were implemented in R using the prop.test function to identify sites differentially unstable in at least one cancer type relative to the average frequency of instability observed across all other groups from the 92,385 sites called in at least half of the samples for each cancer. To determine MSI-H cancer-specific microsatellites, multiple proportions tests were performed for each site for colon, rectal, endometrial, and stomach MSI-H cancers. FDR values were estimated as described above. To compare across cancers and MSI diagnostic types, we computed cosine similarity scores. Because the number of frequently unstable sets in MSI-H cancers was an order of magnitude larger than that observed for MSS cancers, cosine similarity was less sensitive to these set inequalities than the overlap coefficient or Jaccard index, which artificially inflate or deflate the observed overlap, respectively.

Gene Ontology enrichment analyses. Gene enrichment was performed using the R package clusterProfiler version 2.2.5 (ref. 52). clusterProfiler implements a hypergeometric model to test for gene set overrepresentation relative to a background gene set. Each cluster (1–4) from the global instability results was compared with the background of all other microsatellites sequenced at sufficient depth in our study, with a Benjamini–Hochberg FDR threshold of 0.20 defined as significant enrichment. Enrichment in KEGG pathways was analyzed with the enrichKEGG function and the same parameters. Enrichment between MSI-H specific clusters was analyzed using the compareCluster function with fun = enrichGO, pvalueCutoff = 0.05, OrgDb = org.Hs.eg.db. Significantly enriched GO terms were simplified using GOSemSim to calculate the similarity of GO terms and remove highly similar terms (cutoff = 0.7) by retaining the most significant representative term. GO analyses were corrected for gene size in that enrichment analyses were performed at the microsatellite level, such that larger genes required greater numbers of unstable sites for significant enrichment relative to the background distributions of microsatellites in genes covered in our study.

Enrichment of unstable microsatellites in cancer-associated genes. After excluding microsatellites with intergenic and intronic annotations, we extracted annotations for the 252,127 microsatellites that had valid calls in MSS samples, resulting in a panel of 18,104 unique genes. We compared this full gene panel against the 17,564 loci that were unstable with significantly greater frequency in MSI-H cancers at an FDR < 0.05, comprising a set of 6,821 unique genes. We compared these data sets to the COSMIC cancer gene census (accessed 15 June 2015), which contained 573 unique cancer-associated genes. To test for enrichment against the COSMIC database, 1,000 permutations were performed, sampling 6,821 genes from all possible unique genes in the full gene panel.

Correlation of instability and DNA replication timing. We first filtered our data to 77,215 sites that were called in more than half of the samples within each

of the 32 (cancer type by MSI status) groups and called completely across all groups. We downloaded wavelet-smoothed Repli-Seq signals from 11 ENCODE cell lines from the UCSC Genome Browser (GEO GSE34399). We then averaged the proportions of MSI and Repli-Seq signals across 1-Mb windows throughout the genome and calculated the median Repli-Seq signal across all 11 cell lines as representative of 'general' replication timing throughout the genome, with values ranging 0–100 (higher numbers indicating earlier replication). Spearman correlation coefficients were calculated between binned, averaged instability proportions between MSI classifications and across cancer types compared with median and cell-line-specific binned Repli-Seq signals.

Survival analyses. We assessed the association of MSI with overall survival using the coxph function from the R survival package version 2.38, with significance assessed by Wald tests. Age, sex, cancer type, radiation therapy, and pathologic stage (I, II, III, IV) were included as covariates in multivariate analyses. The proportional hazards assumption for covariates in these Cox regression models was tested using the cox.zph function and violating covariates were stratified when necessary.

Statistical analyses. All statistical tests used in this study were nonparametric and therefore made no assumptions about distributions or equal variance between groups. Two-sided Fisher exact tests were used to identify differentially unstable microsatellites in MSI-H cancers and enriched or depleted genomic annotations for unstable sites. To determine unstable microsatellites unique to specific MSI-H cancers, multiple proportions tests were performed for each site across colon, rectal, endometrial, and stomach MSI-H cancers. FDR values for both analyses were estimated using Storey's q-value method, with a q-value < 0.05 considered significant. To compare instability events across cancers and MSI diagnostic types, we computed cosine similarity scores. Hypergeometric tests were implemented to test for the enrichment of genes harboring frequently unstable sites in GO terms and KEGG pathways. Permutation tests were performed to test for enrichment in MSI-affected genes against the COSMIC database. Spearman correlation coefficients were calculated to evaluate correlations between instability and DNA replication timing. Lastly, survival curves were represented with Kaplan-Meier curves, with the significance of covariate effects estimated by fitting Cox proportional-hazards regression models.

Data access. Primary sequencing data are available from TCGA Research Network (http://cancergenome.nih.gov/). Primary MSI calls from this study are available from (http://krishna.gs.washington.edu/content/members/hauser/mosaic/).

Code availability. Code for primary analysis of microsatellite loci through mSINGS (git commit e32b776) is available at https://bitbucket.org/uwlabmed/msings. Code for secondary analyses and MOSAIC are available at https://github.com/ronaldhause/mosaic.

51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

doi:10.1038/nm.4191

Corrigendum: Classification and characterization of microsatellite instability across 18 cancer types

Ronald J Hause, Colin C Pritchard, Jay Shendure & Stephen J Salipante

Nat. Med. **22, 1342–1350 (2016); published online 3 October 2016; corrected after print 19 July 2017**

In the version of this article initially published, in Figure 4d, the column labels UCED and STAD were inadvertently switched. The error has been corrected in the HTML and PDF versions of the article.