A molecular portrait of microsatellite instability across multiple cancers

Isidro Cortes-Ciriano[1,4,*], Sejoon Lee[2,*], Woong-Yang Park[2], Tae-Min Kim[3,†] and Peter J. Park[1,4,†]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

[2] Samsung Genome Institute, Samsung Medical Center, Seoul, South Korea

[3] Department of Medical Informatics and Catholic Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul, South Korea

[4] Ludwig Center at Harvard, Boston, MA 02115, USA

* These authors contributed equally.

[†] Correspondence should be addressed to P.J.P. (peter_park@harvard.edu) and T.M.K. (tmkim@catholic.ac.kr).

ABSTRACT

Microsatellite instability (MSI) refers to the hypermutability of the cancer genome due to impaired DNA mismatch repair. Although MSI has been studied for decades, the large amount of sequencing data now available allows us to examine the molecular fingerprints of MSI in greater detail. Here, we analyze ~8000 exome and ~1000 whole-genome pairs across 23 cancer types. Our pan-cancer analysis reveals that the prevalence of MSI events is highly variable within and across tumor types including some in which MSI is not typically examined. We also identify genes in DNA repair and oncogenic pathways recurrently subject to MSI and uncover non-coding loci that frequently display MSI events. Finally, we propose an exome-based predictive model for the MSI phenotype that achieves high sensitivity and specificity. These results advance our understanding of the genomic drivers and consequences of MSI, and a comprehensive catalog of tumor-type specific MSI loci we have generated enables efficient panel-based MSI testing to identify patients who are likely to benefit from immunotherapy.

Introduction

Microsatellites (MS) are tandem repeats of short DNA sequences, from mononucleotides to longer repeats such as hexanucleotides. MS are abundant throughout the genome and, owing to their high mutation rates, have been widely used as polymorphic genetic markers in population genomics and forensics. Microsatellite instability (MSI) is a hypermutator phenotype that occurs in tumors with impaired DNA mismatch repair (MMR) and is characterized by widespread length polymorphisms of MS repeats due to DNA polymerase slippage[1] as well as by elevated frequency of single-nucleotide mutations (SNVs). MSI is caused by inactivation of MMR genes (e.g., *MLH1*, *MSH2*, *MSH3*, *MSH6* and *PMS2*) through somatic mutations in sporadic cases, with increased risk of cancer for those with inherited germline mutations (i.e., Lynch syndrome)[2]. MSI also occurs by hypermethylation of the *MLH1* promoter (e.g., associated with the somatic *BRAF* V600E mutation)[3], epigenetic inactivation of *MSH2*[4], or down-regulation of MMR genes by microRNAs[5]. MSI events within coding regions can alter the reading frame and lead to truncated, functionally-impaired proteins.[6]

MSI is observed in 15% of sporadic colorectal tumors diagnosed in the United States[7], and has been reported in glioblastomas, lymphomas, stomach, urinary tract, ovarian and endometrial tumors[8]. In clinical settings, detection of MSI is customarily performed by immunohistochemical analysis of MMR proteins or by profiling the Bethesda markers[7], which often include two mononucleotide (BAT25 and BAT26) and three dinucleotide (D5S346, D2S123 and D17S250) MS loci. Colorectal tumors unstable at >40% of the Bethesda markers are considered MSI-High (MSI-H) and are known to have better a prognosis and to be less prone to metastasis than MSI stable (MSS) tumors[9].

It was conjectured more than two decades ago that the less aggressive nature of MSI tumors may be due to their high incidence of somatic mutations, which results in a greater likelihood of having mutated genes whose products elicit antitumor immune responses[10]. Indeed, in melanoma and lung tumors, an elevated mutational load has been associated with an increased rate of response to anti-CTLA-4 and anti-PD-1 therapies, respectively, likely as a result of a higher neo-antigen burden leading to antitumor immune response[11,12]. Other reports have shown that colorectal patients with MMR deficiency have better responses to immunotherapy by PD-1 immune checkpoint blockade and show improved progression-free survival[13]. Although the precise link between the mutator phenotype with MSI and patient response to immunotherapy remains to be elucidated, it is clear that accurate identification of patients with the hypermutator phenotype and their genomic characterization is of therapeutic importance.

In this study, we utilize data from The Cancer Genome Atlas (TCGA)[14] to expand our previous analysis of MSI in 277 colorectal and uterine endometrial exomes to a pan-cancer cohort comprising ~8,000 exomes and ~1000 whole genomes, spanning 23 tumor types. We systematically profile the patterns of MSI mutations in both nuclear and mitochondrial DNA, characterize the affected pathways, and find association with epigenomic features. These analyses uncover new genes harboring frameshift MSI events with varying degrees of cancer-type specificity and generate the most comprehensive catalogue of MS loci selectively subject to DNA slippage events in MSI-H tumors to date. This set includes loci in the non-coding portions of the genome revealed by whole-genome sequencing. Lastly, we describe highly accurate predictive models of MSI-H status based on exome sequencing data.

Results

The exome-wide profiles of MSI in cancer genomes

To obtain an MSI landscape in cancer patients, we analyzed exome sequencing data for 7,919 tumor and matched normal pairs across 23 cancer types (Table 1). We identified 386,396 microsatellite repeats among the 39,496 RefSeq mRNA sequences[15] and tested for the presence of MSI at microsatellites that had sufficient coverage in the exome data (Fig. 1a; Online Methods).

We first investigated five tumor types (COAD: colon adenocarcinomas, ESCA: esophageal carcinoma, READ: rectal adenocarcinoma, STAD: stomach adenocarcinoma and UCEC: uterine corpus endometrial carcinoma) for which the MSI status was determined by TCGA using capillary sequencing-based fragment length assay (Supplementary Table 1)[16–18]. These five tumor types have been recognized as MSI-prone and comprise the majority of MSI events amongst the samples we examined (44,462 MSI events in these five tumor types, *n*=904, versus 29,659 events in the remaining cases, *n*=7,015). Figure 1b shows the abundance of MSI events across the 190 MSI-H cases in these five tumor types (see Supplementary Figs. 1 and 2a for the remaining 118 MSI-L (MSI-Low) and 522 MSS tumor genomes in these 5 tumor types, and Supplementary Fig. 2b and Supplementary Table 1 for all tumor types). Our analysis highlights that MSI mutations represent a continuous rather than a dichotomous phenotype. The figure also shows a pronounced variability in the number of MSI events across MSI-H cases and across cancer types, indicating substantial intra- and inter-tumor type heterogeneity in the genomic impact of MSI (Supplementary Fig. 1 and Table 1). For example, we note that 7% and 17% of the MSI-H tumors have fewer than 10 and 50 detected MSI events, respectively, including one COAD MSI-H tumor without any exonic MSI events, while others have several hundred (exome sequencing data covers some neighboring non-exonic elements including small fractions of UTRs and introns; our term ‘exonic’ includes those spill-over regions).

To discriminate clonal MSI events from subclonal ones, we examined the CCF (cancer cell fraction) using the allele counts of sequencing reads containing the wildtype and mutant alleles[19]. The majority (87%) of the MSI events (32,327 events observed in 146 MSI-H cases for which tumor purity and absolute and local copy number data could be computed[20]) are clonal (CCF > 0.95, Supplementary Fig. 3a). We also observe that the relative abundance of MSI events is largely comparable across tumor stages (I-IV). This suggests that the mutational burden that tumors can tolerate may be similar across early and late tumor stages (Supplementary Fig. 3b). We note that subclonal MSI events may be under-represented relative to clonal events in late stages (Stage II - IV) compared to Stage I (Supplementary Fig. 3b), which would suggest that these genomes are more subject to clonal sweeps during the late stages.

Next, we identified genes with recurrent MSI events using MutSigCV[21]. The genes displaying significant enrichment for coding MSI (FDR < 0.05) along with their level of significance across three tumor types are shown in Suplementary Fig. 4. Pathway analysis reveals that transmembrane/TGFβ, response to cellular stress/DNA damage and chromosome/M-phase-related molecular functions are significantly enriched in genes harboring recurrent MSI in COAD, STAD and UCEC cases, respectively (Supplementary Table 2; P < 0.01).

The mutational landscape of DNA repair pathways

The rates of deleterious mutations (i.e., missense, nonsense and splicing site SNVs, as well as frameshift indels) and frameshift MSI events for *MLH1, MLH3, MSH2, MSH3, MSH6, PMS1, PMS2, POLE, POLD1, PRKDC, APC* and *BRAF* (p.V600E) are shown in Figure 1b. Among these genes, selected on the basis of their known implication in MSI, DNA repair and colorectal cancer, MSI frameshift events represent a major source for the inactivation of *MSH3* (85%) and *MSH6* (85%). In contrast, deleterious SNV mutations more frequently contribute to the loss of function of *POLD1* and *POLE* (27% and 23% of cases, respectively). We next examined the patterns of frameshift MSI events across MSI-prone tumors. We selected a set of 151 genes[22] involved in several DNA repair pathways, including non-homologous end joining (NHEJ), homologous recombination (HR), base excision (BER), RecQ helicase-Like (RECQ), translesion synthesis (TLS), and ataxia telangiectasia mutated (ATM)[22]. We find that COAD samples harboring a large number of MSI events (>500 in our samples) are enriched for *MLH1* promoter hypermethylation (Figure 1b), as previously reported for this tumor type[15]. The genes most frequently targeted by MSI are *RAD50* (16% of MSI-H tumors), *ATR* (15%) and *RBBP8* (10%) (Supplementary Table 3 and Supplementary Fig. 5a).

We have also examined the impact of germline mutations in the MSI-H cases. We observe that 4 COAD (9%), 4 UCEC (5%) and 2 STAD (3%) patients harbor deleterious germline mutations in MMR genes. Of these, at least 5 patients may have acquired the MSI-H phenotypes due to biallelic inactivation of MMR genes, where the inherited germline mutations of MMR genes are complemented with somatically acquired mutations of the corresponding genes. One COAD sample harbored germline and somatic mutations in *MLH1*; 1 STAD and 3 UCEC cases harbor germline and somatic mutations in *MSH6.* Overall, germline mutations in MMR genes, *POLE* and *POLD1* are consistently more prevalent in MSI-H patients compared to MSS cases (Fig. 1c and Supplementary Fig. 6). This frequency of germline mutation carriers in MMR genes may be underestimated since we have applied stringent filtering criteria for our germline calls (Online Methods) to account for the uncertain pathogenicity of missense mutations[23], as well as the technical challenges in identifying mutations in *PMS2* with multiple copies of its pseudogenes in the genome[24].

Although it is difficult to pinpoint the genomic events initiating MMR deficiency, it is likely that truncating mutations in various MMR genes in addition to the hypermethylation of *MLH1* may shape MSI-H genomes, leading to further accumulation of mutations in genes of the DNA repair pathway. To investigate the downstream impact of somatic alterations in MMR genes and proofreading DNA polymerases, we examined the correlation between gene expression and promoter methylation, DNA copy numbers, somatic SNVs and indels and MSI events

(Supplementary Fig. 7). For *MLH1*, only the DNA methylation level is associated with gene expression levels (r = -0.79; Pearson correlation), consistent with a previous report[3]. No apparent relationship between promoter methylation and gene expression is observed for the other genes examined. Other than *MLH1*, the most common genomic events that show association with gene expression (*P* < 0.05; Mann-Whitney test) are the truncating SNVs and frameshift MSI events (*MLH3*, *MSH2*, *MSH3*, *MSH6*, *PMS1* and *POLD1*), suggesting that these somatic events are responsible for the under-expression of the genes in which they are located and may also play a major role in the development of the MSI-H phenotype. The association between DNA copy number and gene expression (r > 0.2; Pearson correlation) is observed for *MSH2* and *POLD1*. We do not observe any significant association between gene expression and germline truncating mutations.

Cancer-type specificity in loci frequently targeted in frameshift MSI events

We investigated the frequency of frameshift MSI events in 130 cancer-related genes[25] across the MSI-prone tumors. Tumor-type specificity of frameshift MSI is evident for some well-known targets of MSI, such as *ACVR2A* (52% of MSI-H tumors) and *TGFBR2* (44%) (enriched in both COAD and STAD; *P* < 0.05, one-tailed Fisher's exact test) as well as *RPL22* (31%), *RNF43* (31%), *MLL3* (27%), *PRDM2* (21%), *JAK1* (16%) and *APC* (3%) (Supplementary Fig. 5b and Supplementary Table 4)[26–28]. For instance, frameshift MSI events are present in *TGFBR2* for 26/45 (58%) of COAD and 51/64 (80%) of STAD but only in 4/75 (5%) of UCEC cases, suggesting that certain tumor types or tumor environments are favorable for the occurrence of particular MSI events. Given the nature of frameshift MSI events in coding regions (i.e.*,* protein-truncating and thus inactivating), the absence of MSI in well-known oncogenes such as *BRAF* likely represents the pressures of negative selection in the context of the MSI-H phenotype. Instead, *BRAF* V600E mutations are observed in 22 of the 45 (49%) MSI-H COAD tumors, but only 4 (2%) frameshift MSI events are observed within the gene.

To uncover other MS loci frequently targeted by frameshift MSI mutations, we first ranked MS loci according to the recurrence level of frameshift MSI events in COAD, STAD and UCEC MSI-H tumors. This analysis resulted in 16,812 frameshift MSI events across a set of 6,441 coding MS loci spanning 4,898 genes, several of which display cancer type specificity (Fig. 2a and Supplementary Table 5). The most recurrent frameshift MSI events are found *in ACVR2A* (51.6% of the tumors)*, KIAA2018 (*51%), *SLC22A9 (*50%), *ASTE1* (45%*), TGFBR2* (44%), *NDUFC2* (36%), *LTN1* (36%) and *SEC31A* (36%). Frameshift MSI events often display significant tumor type specificity, e.g., *MLL3, PRDM2, C9orf114, BAX and OR7E24* are enriched in STAD*, JAK1, TFAM* and *SMC6* are enriched in UCEC, whereas *SEC31A, C18orf34, NDUFC2, KIAA1919* and *CCDC168* among others, are enriched in COAD (P < 0.05, one-tailed Fisher's exact test). Among low-frequency MSI events, *SMAP1*, *CCDC168* and *SPINK5* harbor frameshift mutations in COAD and UCEC but not in STAD tumors.

By analyzing the frequency of MSI events in untranslated regions, we found that MS loci within the 3' UTR region of *C18orf56*, *C14orf169*, *FOXP1*, *UGDH, RNF19B, PUS3* and *FAM60A* as well as the 5' UTR region of *STC1, RBMXL1, RFX1, BEX5* and *SLC6A15* are recurrently altered by MSI across MSI-H cases (Fig. 2b,c and Supplementary Table 6,7). Other MS loci display marked cancer-type specificity, e.g., MSI events within the 5' UTR region of *ZNF738*, *C10orf140, ZNF271 and RAB28* are specific for COAD tumors, whereas *EBP, TMEM182, MIR567* and *MEIS1* are depleted for MSI in STAD tumors. Supplementary Table 8 reports the enrichment of frameshift, 3' and 5' UTR MSI events in COAD, STAD and UCEC.

To obtain a comprehensive MSI landscape on a pan-cancer scale, we next extended our analysis to all exomes, irrespective of their status as MSI-H, MSI-L or MSS. We observe frameshift MSI events in 8,011 MS loci, of which 51 are altered in more than 50 samples (Supplementary Fig. 8 and Supplementary Table 9). *ACVR2A, TGFBR2, KIAA2018, ASTE1* and *SLC22A9* frequently harbor MSI events in STAD and COAD, whereas several other genes are mostly specific to a single tumor type. For instance, *FAM129A, GMIP* and *NEK3* are altered in 107 (12%), 93 (10%) and 53 (6%) BRCA (breast cancer) tumors, respectively; *ALPK2* and *DPYSL2* are altered by frameshift MSI in 73 (17%) and 59 (14%) OV (ovarian serous cystadenocarcinoma) patients, respectively, but only 62 times in the remaining samples; and *ABT1* and *SLC22A24* are altered in 19.6% and 14% of ACC (adrenocortical carcinoma) tumors. Although the cancer-related roles of these novel MSI targets are largely unknown, it has been shown that siRNA-mediated inhibition of *ALPK2* inhibits apoptosis,[29] suggesting that the functional implication of these novel, recurrent MSI events warrants further investigation.

A genome-wide mutational spectrum of MSI

Despite the plethora of studies on genome-wide somatic variation, the landscape of non-coding MSI in cancer remains uncharted territory. Most of the non-coding MSI events are neutral passengers, but those occurring in regulatory elements (see next section) can function as cancer drivers, similar to somatic SNVs in enhancer regions that have been shown to play a role in tumorigenesis.[30,31] To profile the distribution of MSI events genome-wide, we leveraged 708 whole-genome (mean coverage: 55X) pairs of tumor and matched non-neoplastic samples spanning 16 cancer types (Fig. 3 and Supplementary Table 10). The number of MSI events in MSI-H tumors differs significantly from that in MSI-L ($P = 6.25 \times 10^{-11}$, Kolmogorov-Smirnov) and MSS ($P = 4.01 \times 10^{-15}$, Kolmogorov-Smirnov) tumors, whereas the numbers are comparable between MSI-L and MSS cases ($P = 0.17$, Kolmogorov-Smirnov). As shown in Fig. 3a, when samples are ordered by decreasing number of MSI events within each tumor type, the decrease is gradual, confirming that MSI represents a continuous rather than a dichotomous phenotype.

The numbers of exonic MSI calls between the exome and whole-genome datasets show a high correlation (r = 0.90, $P < 10^{-15}$, Pearson correlation for 531 cases profiled on both platforms; Supplementary Fig. 9). However, many MSI events are missed in whole-genome data due to their lower coverage, with only 32% of the calls from the exomes captured on the genomes (based on 23 pairs with at least 50 MSI events in exome data). On the other hand, 13% of the exonic MSI events identified in whole-genome data are missed in the exome-based calls, since the target capture regions do not include many exonic MS repeats.

To further analyze the influence of read depth on MSI detection sensitivity, we performed subsampling analysis, using a sample with a high number of MSI events and coverage (TCGA-AD-A5EJ; tumor at 82X, matched normal at 42X) (Supplementary Figs. 10-11). We find that the number of MSI events recovered decreases substantially when the coverage is reduced to 20-30X. However, we do not see any clear relationship between the number of MSI events and coverage above that range (Supplementary Fig. 11), suggesting that sequencing coverage is not a major factor, especially given that most MSI events are clonal as we described earlier. We also examine the correlation between the number of MSI events and tumor purity (Supplementary Fig. 12).[20] The number of MSI events identified in high purity samples spans five orders of magnitude, whereas we systematically detect fewer than a thousand MSI events in the case of low-purity samples. However, the number of low purity samples (e.g., <0.6) in our set is small, meaning that the impact of coverage due to variation in purity does not have a substantial impact on our analysis.

The genome-wide density of MSI events along the chromosomes does not show statistically significant correlation with SNV density, regardless of the bin size (100kb - 10Mb bins;

Supplementary Fig. 13). We have previously used exome sequencing to identify a portion of MSI events in UTRs[15]. With whole-genome data, the number of MSI calls in 3’ UTRs is substantially larger (Supplementary Fig. S9b-d). This allowed us to find that MSI events are enriched in 3’ UTRs for MSI events in 21 out of 25 MSI-H cases (84%), whereas they are depleted in 5’ UTRs and coding regions in 24 out of 25 MSI-H cases. In MSS tumors, only 3 out of 105 (3%) show enrichment of MSI in 3’ UTR regions, whereas 42 (46%) show depletion of MSI events in coding regions ($P < 0.05$; one-tailed Fisher’s exact test). Overall, these results suggest that MSI events in 3’ UTRs may be under positive selection in MSI-H tumors. The shortening of 3' UTRs in cancer cells is known to increase the stability of transcripts and thus the translational level of oncogenes[32]. The frequent MSI events in 3' UTRs may have similar functional consequences (e.g., the loss of miRNA-mediated regulation), although they often result in down-regulation of their corresponding genes. [33]

We find that 62% of the tumors harbored more than 100 MSI events genome-wide, including samples from all 16 cancer types examined. We again observe a substantial level of intra-tumor type heterogeneity in MSI abundance, with the number of MSI events varying up to 5 orders of magnitude. The presence of noncoding MSI events genome-wide in tumor types beyond the MSI-prone cases is notable. For instance, we observe that 81% and 53% of OV and KICH (kidney chromophobe) samples, harbor >1,000 MSI events, exceeding the previously reported MSI-H frequencies of OV (12%) and overall kidney cancers (6%)[34,35]. Elevated microsatellite alterations at selected tetranucleotides (EMAST) has been observed in non-MSI-prone tumor types such as lung, head and neck cancers as well as melanoma[36], but the current markers for EMAST could not be captured with the short reads in our data due to their size.

To investigate the relationship between epigenetic features and the genome-wide distribution of MSI events, we selected the 50 genomes displaying the highest MSI rates and compared their MSI density with the 25-state chromatin state map based on 12 epigenetic marks across 127 epigenomes[37] (Online Methods; Supplementary Figs. 14-17). For best estimates, the chromatin state map from the most ‘similar’ tissue type was used for each tumor type. Our analysis reveals significant enrichment for MSI events in actively-transcribed regions, promoters and enhancers in most MSI-H genomes (two-tailed Fisher’s exact test, $P < 0.05$; Supplementary Fig. 18). On the other hand, inactive regions, including constitutive heterochromatin, repressed Polycomb, bivalent promoters and quiescent chromatin, are significantly depleted for MSI overall. Taken together, these results show the over-representation of MSI in functionally important, typically open-chromatin regions, extending our previous results based on seven colorectal and UCEC tumors[15].

MSI events in the mitochondrial genome

To obtain a mitochondrial MSI landscape, we analyzed TCGA low-coverage (6-8X) whole-genome sequencing data that we have generated (the sample size in this dataset was larger than that of the high coverage dataset). Due to their high copy number, the mitochondrial DNA had a median coverage of >800X even in the low-coverage data[38]. We applied our MSI discovery pipeline to a set of 31 mitochondrial MS loci (Supplementary Table 11) across 308 cancer genomes (COAD, STAD and UCEC) (Fig. 3b and Supplementary Table 12). The most recurrent MSI event is observed in *DQ582201*, a polyC mononucleotide repeat (115 MSI events, 37% of tumors); the second most recurrent event is on the exon regions of *AF079515* (15 MSI events, 5% of tumors). The instability of *DQ582201* has been reported in several cancer types.[39] The majority of mitochondrial MS loci (22 out of 31 MS loci; 71%), however, do not contain MSI events in any of the tumors examined, suggesting that mitochondrial MSI is not widespread compared to nuclear MSI. Moreover, mitochondrial MSI events are not associated with the MSI status of the tumor: 43 (54%), 22 (48%) and 98 (54%) mitochondrial MSI events are observed in MSI-H, MSI-L and MSS genomes examined, respectively. The relationship between the nuclear MSI and mitochondrial MSI has been controversial[39], but our mitochondrial genome-wide MSI examination suggests that these two events are not correlated.

Defining a panel of hypermutable MS loci enriched in MSI-H cases

We focused on frameshift MSI events earlier to understand the functional impact of coding MSI in MSI-prone tumors. Here, we sought to uncover mutational patterns discriminative of MSI-H status across exonic and noncoding MS loci, as well as to identify hyper-mutable MS loci across the entire genome. We first ranked recurrently-targeted MS loci by their specificity for MSI-H tumors using exome sequencing data from COAD, STAD and UCEC tumors. Our analysis yielded a catalogue of MS loci specific to MSI-H tumors (Fig. 4a and Supplementary Table 13). We find that several of these MS loci lie within genes prone to frameshift MSI events, such as *KIAA2018*, *ACVR2A* or *ASTE1*[15]. In contrast to frameshift and 3'/5' UTR MSI events (Fig. 2), few MSI events enriched in MSI-H and depleted in MSS/MSI-L cases display cancer-type specificity, implying that there are commonalities in the molecular mechanisms underlying MSI at these loci across the three cancer types.

Given that most MS loci lie within the non-coding genome, we also extended our recurrence analysis to whole-genomes by utilizing sequencing data from 25 MSI-H, 19 MSI-L and 105 MSS tumors. We discovered a set of intronic and intergenic MS repeats recurrently targeted by MSI in MSI-H cases (Fig. 4b and Supplementary Table 14). Perhaps not surprisingly given the larger list of candidate MS, these non-coding loci are more specific to MSI-H than the best exonic loci are (cf. Fig. 4a), containing MSI events in nearly all of the MSI-H tumors and almost none in MSI-L or MSS samples. These inquiries have yielded a collection of coding and non-coding MS loci recurrently targeted by MSI in MSI-H tumors, which provide a foundation to refine and extend the set of markers employed for MSI-H categorization.

Prediction of MSI status from exome-sequencing data

The total numbers of MSI and frameshift MSI events are significantly higher in MSI-H tumors than in MSI-L or MSS tumors ($P < 10^{-15}$; Kolmogorov-Smirnov test; Figs. 5a,b). The number of SNV and MSI events exhibit moderate to low correlation in MSI-H (r = 0.32, $P = 6.15 \times 10^{-6}$ in exomes, and r = 0.35, P =0.09 in whole genomes; Pearson correlation; Figs. 5c,d), MSI-L (r = 0.10, P = 0.68, Pearson correlation) and MSS (r = -0.06, P = 0.56, Pearson correlation) tumors. To test whether our MSI calls can be used to distinguish between MSI-H and MSS cases, we built Random Forest[40] classification models. Each tumor was encoded by a vector recording the presence or absence of MSI events across MS loci, as well as the total number of MSI events (Online Methods). Models built on a limited set of learning examples (i.e., only those 190 MS-H, 118 MSI-L and 533 MSS tumors with MSI status annotations) are likely to possess limited predictive power on external data. Thus, we included conformal prediction[41] in our modeling pipeline to provide confidence estimates for individual predictions (Fig. 5e,f; Online Methods). Briefly, conformal prediction evaluates the similarity (i.e., conformance) between the new samples and the training data. The output represents the probability that the new sample is either MSI-H, MSS, or uncertain (in the case of the new samples being outside the applicability domain of the model), given a user-defined tolerance level, which sets the maximum allowable fraction of erroneous predictions. Our 10-fold cross-validation (CV) showed high accuracy of the models produced (sensitivity: 92%; specificity: 99%). Comparable results were obtained in Leave-One-Out CV (sensitivity: 93%; specificity: 99%), indicating that the MSI events detected using whole-exome data convey enough predictive signal for MSI categorization.

By applying the prediction model to 7,089 exomes from 17 cancer types not commonly tested for MSI status, we identified 91 additional MSI-H cases using a confidence level of 0.75, 22 of which were identified at confidence level of 0.80 (Figs. 5g,h and Supplementary Table 15). Among the 91 cases, the most frequent are BRCA (16), OV (14) and LIHC (liver hepatocellular carcinoma ; 11). Our estimated MSI-H rate for ovarian cancer is 3.2%, significantly lower than that reported previously (10%)[42]; for HNSC (head and neck squamous cell carcinoma) and CESC (cervical cancer), our estimated MSI-H rates are 1.2% and 2.3% whereas the reported rates in the literature are 3% and 7%.[8] The frequencies generated for the other non-MSI-prone cancer types were mostly in agreement with the reported numbers in the literature.[8] For example, our estimated MSI-H frequencies for PRAD (prostate adenocarcinoma), LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma) are 0.6%, 0.2% and 1.2%, respectively, which are comparable to the frequencies of 1% and 0-2% reported for prostate and for lung cancers, respectively[8]. We note that the differences in the rates may be due to the small sample sizes used in the literature for some tumor types[8], differences in the characteristics of the cohorts (e.g., tumor stage) and tumor-type specific features that were missed in our model. We did not identify any MSI-H cases among THCA (papillary thyroid carcinoma; *n*=493), PHCA (pheochromocytoma; *n*=179) and SKCM (skin cutaneous melanoma; *n*=109) tumors. Overall, the frequency of MSI-H cases in non-MSI-prone cancer types was found to be 1.3%, significantly lower than the 14% we observed in UCEC, STAD, COAD, READ and ESCA tumors. Consistent with our analyses of COAD, READ, STAD, ESCA and UCEC MSI-H tumors (Fig. 1b), we found that the number of MSI events varied markedly across these newly identified MSI-H tumors (Fig. 5h). We detected 1,365 frameshift MSI events in the tumors predicted as MSI-H, with the most frequent incidences in *DPYSL2* (12 cases), *OR11G2* (9), *SLC22A9* (9) and *KIAA2018* (8), suggesting that the MSI events that recur in MSI-H cases (*cf.* Figure 2) constitute a mutational signature that is leveraged by the predictive model for MSI categorization. We find that 31 patients display somatic mutations in MMR genes, and 1 CESC (TCGA−EA−A410) and 2 LIHC (TCGA-WQ-A9G7 and TCGA-EP-A12J) cases harbor germline mutations in *MSH2*, *MSH6* and *MLH3*, respectively. Additionally, we observe that 1 BRCA patient (TCGA-BH-A18G) harbors a missense germline mutation predicted to be pathogenic with high confidence (Online Methods) and a somatic frameshift event in *MSH3*.

We also performed mutation signature analysis based on the mutation frequencies of 96 trinucleotide contexts[43] (Supplementary Fig. 19). For the 91 MSI-H predicted cases, we confirm the mutation signatures characteristic of MSI-H cases, e.g.*,* C>T transitions in (A/C/G)pCpN sequence contexts and C>A transversions at an CpCpN context, suggesting that the mutation signatures of predicted MSI-H cases are largely concordant with those of known MSI-H cases.

Discussion

Our pan-cancer profiling of exome and whole-genome sequencing data for MSI has demonstrated that the number of MSI events varies profoundly within and across tumor types. Moreover, this variability is also observed within the MSI categories established by the Bethesda guidelines (MSI-H, MSI-L and MSS). These data therefore support the argument that microsatellite instability represents a continuous phenotype.

A joint analysis of MSI-H tumors from multiple cancer types has revealed that several DNA repair pathways other than MMR, including ATR, BER, HR and NHEJ, are altered by single-nucleotide and MS mutations. Moreover, we have uncovered new genes affected by frameshift MSI events in MSI-prone tumors as well as in tumor types not frequently affected by MSI (e.g.*, FAM129A, GMIP* and *NEK3* in BRCA, and *DPYSL2 and ALPK2* in OV). Some of these genes have shown strong predictive power for MSI-H status (e.g.*, ACVR2A* and *KIAA2018* for COAD, STAD and UCEC), whereas others display low recurrence and specificity for single cancer types (e.g.*, SMAP1* for STAD) Along with the diverse molecular functions enriched for MSI events in these tumor types, our data reaffirm that some genes are particularly susceptible to MSI in certain cancer types[44]. Although some of their potential cancer-related roles have been identified[45], the functional relationship between MSI and tumorigenesis as well as the similarity of molecular mechanisms establishing MSI phenotype across cancer types remain to be validated.

By classifying 7,089 patients into MSI-H and MSS categories using our MSI-based predictive model, we identified 91 new MSI-H cases from 16 different tumor types. According to our classification model, the frequency of MSI-H cases in MSI-prone tumors is roughly ten times larger than in other tumor types (14.5% vs 1.3%). In contrast to previous models based on SNVs or mononucleotide repeats,[46,47] our modeling approach estimates the likelihood of prediction error for individual patients using a confidence level, which can be easily interpreted (e.g., a confidence level of 0.9 means that no less than 90% of the predictions should be correct).

Although the search space considered in our whole-genome and exome MSI reference sets is large (~19 million and 386,396 MS loci, respectively), it comprises only MS repeats of size 6-60 bp and up to tetrameric repeats. Although we are missing longer repeats as well as MS loci longer than 60 bp, our MSI calling pipeline captures the vast majority of MS loci (e.g., >99% of repeats in the final set of exome and genome reference MS are smaller than 40 bp).[15] Nonetheless, MSI events in certain non-coding MS loci might have missed the threshold for significance due to low coverage, and we anticipate that the rates of MSI events presented here are likely to be under-estimates of the true rates. The sequencing data used in this study were obtained from a sequencing platform (Illumina) that is known to be robust in estimating the length of homopolymer runs,[48] but a platform with longer reads will help in a more comprehensive identification of MSI target loci. A further increase in sample size, especially those with annotated MSI status to include in a better training set, will also increase the power to detect all relevant loci.

Of clinical relevance, we provide the largest catalogue available to date of coding and non-coding MS loci frequently altered across cancers. This repertoire has implications for MSI categorization and provides the groundwork for further investigation into the molecular mechanisms underlying MSI recurrence. For many cancer patients, the use of targeted panels containing only a subset of cancer-related genes is informative for personalized therapeutics. Such panels offer the advantage of very high sequencing depth (e.g., >1000X) to detect somatic mutations with low variant allele frequencies that would be missed by standard exome sequencing. The loci identified in our study, especially those in the non-coding regions, can be included in panel-based assays to serve as highly sensitive markers for MSI across multiple tumor types. This will avoid a separate test for MSI in MSI-prone tumors, and it will identify a small set of MSI patients in non-MSI-prone tumors in which MSI is rarely considered by clinicians. The potential benefit of such a panel is enormous, given the demonstrated efficacy of immunotherapy for the MSI cohort.[13]

Online Methods

Data sets

Exome and whole-genome tumor-normal pairs from the TCGA project were downloaded from CGHub (http://cghub.ucsc.edu). The reads were mostly 100 bp paired-end reads and were aligned to the NCBI build 37 (hg19) using BWA. The full list of samples is given in Supplementary Tables 1, 10 and 12. The MSI status (MSI-H, MSI-L and MSS) were downloaded from the GDAC (https://gdac.broadinstitute.org) website, whereas the methylation state of the *MLH1* promoter, gene expression and DNA copy number variation data were downloaded from the Genomics Data Commons Data Portal (https://gdc-portal.nci.nih.gov/) websites. MSI status was evaluated by the TCGA consortium for COAD, READ, ESCA, STAD and UCEC tumors using a panel of 4 mononucleotide repeats (BAT25, BAT26, BAT40 and TGFBRII) and 3 dinucleotide repeats (D2S123, D5S346 and D17S250), except for a subset of COAD/READ genomes evaluated by 5 mononucleotide markers (BAT25, BAT26, NR21, NR24 and MONO27)[49]. Tumors were classified as MSI-H (>40% of markers altered), MSI-L (<40% of markers altered) and MSS (no marker altered).

Defining a reference set of MS repeats

To generate an exome-wide reference set of MS loci, we utilized the Sputnik algorithm[15] to identify MS loci in the mRNA sequences of 39,496 RefSeq genes (USCS Genome Browser, hg19). We limited our analysis to mono-, di-, tri- and tetranucleotide MS loci of size 6-60 bp, which can be detected reliably with enough flanking sequences from 100 bp reads. We derived the reference set of MS repeats from RefSeq sequences as the target regions used by the TCGA are different across cancer types. MS repeats falling within splice sites were removed, as they have undetermined genomic coordinates or are redundant to multiple isoforms. The final reference set of exonic MS sites (Supplementary Table 16) comprised 386,396 loci (112,896 mono-, 63,162 di-, 132,117 tri- and 78,221 tetranucleotides). These included 154,590 coding, 50,598 5'-UTR and 181,208 3'-UTR MS loci, as annotated in the UCSC Genome Browser.

To produce a genome-wide reference set of MS loci (Supplementary Table 17), a total of 19,039,443 MS repeats were obtained using the Sputnik algorithm (chromosome 1 through Y) and categorized into five groups (coding, 248,100; 5'-UTR, 39,582; 3'-UTR, 166,111; intronic, 8,265,436; intergenic, 10,320,214). This MS set encompasses 7,404,614 mono-, 3,686,129, di-, 3,750,887 tri- and 4,197,813 tetranucleotides. We also utilized the Sputnik algorithm to build a reference set of mitochondrial MS loci from the hg19 mitochondrial DNA (mtDNA), which contained a total of 31 MS loci (10 mono-, 2 di-, 11 tri- and 8 tetranucleotides) (Supplementary Table 11).

Detection of DNA slippage events

After filtering reads with low mapping quality, intra-read MS repeats were identified with the same method used to identify reference MS repeats and were intersected with the reference MS repeats. We note that the minimum size of intraread repeats detected was 5bp. Thus, reads spanning MS repeats contracted below 5bp were not considered. We required the 2 bp flanking sequences (both 5' and 3') of the intra-read MS repeats to be identical to those of matching reference repeats, thereby discounting truncated MS repeats. In each genome, the distribution of the allelic repeat length at each MS locus was obtained by collecting the lengths of all intra-read MS repeats mapped to that locus. We compared the distributions of MS lengths from tumor and matched normal genomes at each locus using the Kolmogorov-Smirnov statistic. An FDR (false discovery rate) of < 0.05 was used as a threshold for statistical significance, with a minimum of 5 tumor and 5 matched normal reads. We note that the number of MSI ''events'' refers to the absolute number of MSI counts per sample, whereas sample percentage refers to the percentage of samples from a given cancer type harboring MSI events at a particular MS locus. We distinguished MSI events at coding sequences into in-frame and frameshift events depending on whether the difference between (i) the mode of the read length distribution of the normal samples and (ii) the mode of the read length distribution of the tumor sample or the second most frequent read length from this distribution (if supported by at least 20% of the reads) was a multiple of three.

Mutation calling

We utilized MuTect 1.1.4[50] to call somatic mutations in both the tumor and matched normal whole-genome samples, using the Catalogue of Somatic Mutations in Cancer (COSMIC) v68 and dbSNP135 as reference sets of known somatic and germline mutations, respectively. To ensure the somatic origin of the variant sets reported by MuTect, we filtered out germline mutations from the 1000 Genomes Project (phase 3, release 20130502)[51] and any mutation present in at least one read in two unmatched normal BAM files from the same tissue. Somatic mutations for all 7,919 exomes were downloaded from the GDAC (https://gdac.broadinstitute.org) website. We utilized HaplotypeCaller 3.4-46-gbc02625[52]to examine germline mutations. We only kept deleterious mutations (i.e.*,* frameshift, nonsense, missense and splicing site) supported by at least 10 reads, and those with at least 30% of the reads mapped to that locus supporting the alternative allele. Additionally, we only kept missense mutations with a predicted MetaLR score from Annovar[53] higher than 0.9. We did not consider mutations in the exons 9, and 11 to 15 of *PMS2*, as the *PMS2CL* pseudogene displays more than 98% sequence identity with these exons. Due to the high allelic diversity of *PMS2CL* due to sequence transfer[24]*,* it is challenging to dismiss false positive mutations called in these exons.

Correlation between gene expression and genomic events in MMR genes and proofreading polymerases

To investigate the association of the level of gene expression and genomic events on seven MMR genes (*MLH1, MLH3, MSH2, MSH3, MSH6, PMS1* and *PMS2*) and two proofreading DNA polymerases (*POLD1* and *POLE*), we utilized gene expression, promoter methylation and DNA copy number profiles for the186 MSI-H cases with available from these three data types. Gene expression profiles were first log transformed, i.e.*;* $\log_2$ (FPKM+1). Subsequently, the expression values of each row and column were median-centered and rescaled so that the sum of the squares of the values to be 1.0. To process the promoter methylation data, we collected 17 common probes corresponding to the 9 genes studied between two microarray platforms (humanmethylation27 and humanmethylation450, Illumina). β values were obtained for 17 probes and averaged per gene. The *MLH1* promoter was considered methylated in samples with β values > 0.3. To obtain copy number data, we selected segmentation files filtered for germline alterations. Log2 copy numbers overlapping the genomic segments of 8 genes were considered as the copy numbers of these genes. *POLE* was ignored since it was not covered by the segmentation files. Pearson's correlation was used to assess the relationship between gene expression and promoter methylation (β values), as well as the relationship between gene expression and DNA copy numbers. The relationship between gene expression and somatic mutations and MSI events, was evaluated using the Mann-Whitney test (α = 0.05).

Analysis of epigenomic features

We downloaded the coordinates of the 25-state chromatin state map defined using 12 marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H4K20me1, H3K79me2, H3K36me3, H3K9me3, H3K27me3, H2A.Z and DNase) across 127 reference epigenomes from the

Epigenome Roadmap project[54]. For each of the 30 whole genomes with the highest MSI counts, the list of their MSI loci was intersected with the chromatin state maps defined using cell lines from the same anatomical location as the tumor types. We used the chromatin state maps defined using the epigenomes E092, E094, E0110 and E0111 for STAD, E117 for UCEC, E076, E106 and E075 for COAD, E086 for KICH, E027, E028 and E199 for BRCA, E053, E054, E067, E068, E069, E070, E071, E072, E073, E074, E081, E082 and E0125 for GBM, E097 for OV, E088, E096, E114 and E128 for LUSC, E055, E056, E057, E059, E061, E126, E127 and E058 for HNSC and E086 for KIRP. Subsequently, the percentage of MSI events overlapping each chromatin state was averaged across the matched epigenomes. The same process was applied to the set of MS loci from the genome-wide reference set. Fisher's exact test was used to assess the significance of the enrichment for MSI events of each chromatin state in each of the cancer genomes. The significance level was set to 0.05.

MSI status prediction

We used Random Forest models[40] to build binary classifiers for the prediction of MSI status. Each tumor was encoded with a vector recording the number of MSI events and the presence or absence of MSI events in 7,863 genes targeted by MSI in at least one sample. Features displaying a variance close to zero across all learning examples (i.e., near-zero variance descriptors) were removed using the function *nearZeroVar* from the R package *caret*[55]. The remaining descriptors were mean-centered to zero and scaled to unit variance to obtain *z*-scores using the function *PreProcess* from the same package.

The final set of covariates, the MSI status of the training data and the code used to train the models are given in the Supplementary File 1-2. The number of trees was set to 100[41], the optimal value of the parameter *mtry* was determined to be 182 *via* 10-fold cross-validation and the default values were used for the remaining parameters. With this *mtry* value, the final prediction models were built using all available learning examples.

To estimate prediction errors, we used the following pipeline[41] from the R package *conformal* (https://cran.r-project.org/web/packages/conformal/index.html). We used cross-validation predictions to define a Mondrian class list for each category (i.e., MSI-H and MSS) by sorting in increasing order the fraction of trees voting for that class for each training example. Next, we applied the model trained on all learning examples to each sample without MSI categorization, and calculated for each case the fraction of trees in the forest voting for each class. These values were intersected with the corresponding Mondrian class list. For each sample, the *P* value for a given class was calculated as the number of elements in the corresponding Mondrian class list higher than the vote fraction for that class divided by the number of elements in that list. If the *P* value for a given class is above the tolerated error, ε, the sample is predicted to belong to that category. Hence, a given sample may be called as MSI-H or MSS. However, it can also be called as both in cases when the model does not have enough predictive power to discriminate between classes, or neither in cases when the sample is outside the applicability domain of the model. This flexibility thus gives an unbiased estimate of the reliability of the predictions given the training data. The tolerated error, ε, indicates the maximum fraction of predictions that are incorrect. Therefore, increasing the confidence level might increase the number of uncertain predictions, that is, samples classified as both MSI-H and MSS.

Acknowledgements

The results published here are based upon data generated by The Cancer Genome Atlas and obtained from the Database of Genotypes and Phenotypes (dbGaP) with accession number phs000178.v8.p7. Information about TCGA can be found at http://cancergenome.nih.gov. We thank the Research Information Technology Group at Harvard Medical School and KT for providing computational resources. This work was supported by grants from the Ludwig Center at Harvard (I.C.C. and P.J.P.) and the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HI13C2096, S.L. and W.-Y.P.). We thank Alison Barton and Lixing Yang for their critical reading of the paper.

Author contributions

S.L. and I.C.C. performed bioinformatic analysis of all data, with guidance from T.M. and P.J.P. S.L. was supervised by W.-Y.P.; I.C.C. was supervised by P.J.P. The manuscript was written by I.C.C. and P.J.P. with substantial input from T.K. and S.L.

Figures

Figure 1 Schematic overview of the MSI calling pipeline. (**a**) A reference set of exonic and genome-wide MS repeats was assembled from the human reference genome hg19. The sequencing reads spanning each MS repeat and at least 2 base pairs at each flanking side were extracted from the tumor and normal BAM files. This process was repeated for all MS repeats in the reference sets across all pairs of matched normal-tumor samples. The Kolmogorov-Smirnov test was used to evaluate whether the read length distributions from the normal and tumor samples differed significantly (False Discovery Rate < 0.05). The exonic and genome-wide MSI calls served to identify MS loci recurrently altered by MSI in MSI-H tumors, discover frequent frameshift mutations and to predict MSI status. (**b**) Landscape of somatic MSI in MSI-H tumors**.** MSI events (frameshift and in-frame), deleterious SNV (missense, nonsense, and splice site) and indel (frameshift) rates in 190 MSI-H exomes. Samples harboring hypermethylation of the *MLH1* promoter are denoted by blue squares. Deleterious germline and somatic mutations (i.e.*,* missense, nonsense, splice site and frameshift) are depicted in black and red, respectively, whereas frameshfit MSI events are shown in green. Black arrows mark patients with germline and somatic mutations in MMR genes. (**c**) Germline and somatic mutations in MMR genes, *POLE* and *POLD1* in MSS, MSI-L and MSI-H tumors. The heatmap and the cell labels report the number and percentage of samples in each category harboring mutations, respectively.

Figure 2 MS loci recurrently altered by MSI. (**a**) Coding MSI loci recurrently targeted by frameshift MSI in COAD, STAD and UCEC MSI-H tumors. The heatmap shows the fraction of COAD, STAD and UCEC MSI-H tumors containing frameshift MSI events in MS loci located within the coding sequence of the genes indicated on the *x*-axis. The total count of frameshift MSI events at these loci is depicted in the above barplot. The full list of MS loci recurrently altered by frameshift MSI is given in Supplementary Table S4. Similarly shown for genes with frequent 3' UTR (**b**) and 5’ UTR (**c**) MSI events in 3 MSI-prone tumor types.

Figure 3 Pan-cancer landscape of genome-wide MSI. (**a**) The first panel shows the number of MSI events across 708 whole genomes, stratified by the length of the repeat unit. The second and third panels report the MSI status and the total count of SNVs, respectively. The fourth panel shows the distribution of MSI events across the genome. (**b**) Landscape of MSI in

mitochondrial DNA across 308 COAD, STAD and UCEC low-pass whole genomes. MSI events, including frameshift and in-frame mutations, are shown in red.

Figure 4 MS repeats recurrently altered by MSI in MSI-H tumors. (**a**) The barplots report the number of COAD, STAD and UCEC tumors harboring MSI events at the loci indicated in the central panel. This analysis examined 190 MSI-H, 118 MSI-L and 522 MSS exomes. (**b**) The recurrence analysis was extended to 25 MSI-H, 19 MSI-L and 105 MSS whole-genomes. Genomic coordinates in (**a**) and (**b**) indicate the location of the MSI repeats in the hg19 assembly of the human genome.

Figure 5 Distribution of the number of MSI and prediction of MSI status. Distribution of the number of MSI (**a**) and frameshift MSI events (**b**) in MSI-H and MSS (also including MSI-L) tumors. Correlation between the number of SNV and MSI events in exomes (**c**), and whole genomes (**d**). Prediction of MSI status from exome-sequencing data using conformal prediction and Random Forest models (**e**). Initially, we used 5-fold cross-validation to calculate predictions for all training examples. The fraction of trees in the forest voting for each class was recorded, and subsequently sorted in increasing order to define one Mondrian class list *per* category. (**f**) The model which was trained on all training data was applied to 7,089 exomes. For each of these samples, the algorithm recorded the fraction of trees voting for each class. The *P* value for each class was calculated as the number of elements in the corresponding Mondrian class list higher than the vote for that class (e.g. 6 out of 7 in the toy example depicted in Fig. 5a) divided by the number of elements in that list. If the *P* value for a given class is above the tolerated error, ε, the sample is predicted to belong to that category. The confidence level (1-ε) indicates the minimum fraction of predictions that are correct. (**g**) Number of samples predicted as MSI-H, MSS and uncertain (both: in cases when the classifier does not have enough power to confidently assign a single category; none: in cases when the samples that are outside the applicability domain of the model). Here, the confidence level was set to 0.75. (**h**) Landscape of MSI for the 91 exomes predicted as MSI-H at a confidence level of 0.75. Samples predicted to be MSI-H at a confidence level of 0.80 are marked with black arrows.

Supplementary Figures

Figure S1 Distribution of MSI events in MSI-H, MSI-L and MSS cases in MSI-prone tumor types.

Figure S2 (**a**) Landscape of MSI in MSI-L and MSS tumors**.** (**b**) Distribution of MSI events in low-frequency MSI-H tumor types. Blue points indicate the median and the interquartile range (25th through 75th percentile), whereas red points indicate the mean value.

Figure S3 The abundance of clonal and subclonal MSI events is shown for 146 MSI-H cases (**a**). Distribution of clonal and subclonal MSI events across tumor stages.

Figure S4 Top-ranked genes with significantly recurrent MSI in STAD (**a**), UCEC (**b**) and COAD (**c**) tumors (Q value, FDR < 0.05; MutSigCV).

Figure S5 Abundance of frameshift MSI events in genes implicated in DNA repair (**a**) and tumorigenesis (**b**) across 5 MSI-prone cancer types.

Figure S6 Frequency of SNVs and indels in MMR genes, *POLE* and *POLD1* in COAD (**a,b**), UCEC (**c,d**) and STAD (**e,f**).

Figure S7 Gene expression, promoter methylation and DNA copy number profiles for MMR genes (*MLH1, MLH3, MSH2, MSH3, MSH6, PMS1* and *PMS2*) and two proofreading DNA polymerase (*POLD1* and *POLE*) corresponding to186 MSI-H cases.

Figure S8 Coding MSI loci recurrently subject to frameshift MSI events in the 22 cancer types studied (Table 1). This analysis included MSI-H, MSI-L and MSS tumors. The heatmap shows the fraction of tumors from each cancer type harboring frameshift MSI mutations in MS loci located within the coding sequence of the genes indicated on the *x* axis. The total count of frameshift MSI events at those MSI loci across all tumor types is depicted in the above barplot.

Figure S9 Cross-platform comparison of MSI calls in exonic (coding and 3'/5' UTR) (**a**), 3'UTR (**b**), 5'UTR (**c)** and coding (**d**) MS repeats.

Figure S10 Influence of read depth on MSI calling across increasing larger random subsamples of reads from the tumor (82X) and matched normal (42X) bam files corresponding to the patient TCGA-AD-A5EJ. The total number of MSI events and the percentage of MSI called at different subsampling levels are depicted in (**a**) and (**b**), respectively.

Figure S11 Average read depth of the tumor (y-axis) and matched normal (x-axis) samples.

Figure S12 Correlation between MSI rates *vs* tumor purity.

Figure S13 Distribution of Pearson correlation coefficients for SNV and MSI rates measured in (**a**) 1Mb, (**b**) 10Mb and (**c**) 100Kb bins. Blue points indicate the median and the interquartile range (25th-75th percentile), whereas red points indicate the mean value.

Figure S14 Frequency of MSI events in the 25 states of the chromatin state map averaged across 127 epigenomes for samples TCGA-06-0152 (**a**), TCGA-24-1557 (**b**), TCGA-A5-A0G9

(**c**), TCGA-09-2050 (**d**), TCGA-24-1614 (**e**), TCGA-A5-A0GA (**f**), TCGA-13-1491 (**g**), TCGA-34-2600 (**h**), TCGA-A6-6781 (**i**), TCGA-14-1034 (**j**), TCGA-37-4135 (**k**) and TCGA-AA-3518 (**l**).

Figure S15 Frequency of MSI events in the 25 states of the chromatin state map averaged across 127 epigenomes for samples TCGA-AA-A00R (**a**), TCGA-AD-A5EJ (**b**), TCGA-AP-A0L8 (**c**), TCGA-AA-A01R (**d**), TCGA-AO-A124 (**e**), TCGA-AP-A0L9 (**f**), TCGA-AC-A2BK (**g**), TCGA-AP-A051 (**h**), TCGA-AP-A0LD (**i**), TCGA-AD-6964 (**j**), TCGA-AP-A054 (**k**) and TCGA-AP-A0LE (**l**).

Figure S16 Frequency of MSI events in the 25 states of the chromatin state map averaged across 127 epigenomes for samples TCGA-AX-A05S (**a**), TCGA-B6-A0I6 (**b**), TCGA-CG-4442 (**c**), TCGA-AX-A0J1 (**d**), TCGA-BR-4280 (**e**), TCGA-CG-5723 (**f**), TCGA-B5-A11G (**g**), TCGA-BR-6452 (**h**), TCGA-CN-6011 (**i**), TCGA-B5-A11H (**j**), TCGA-BS-A0TE (**k**) and TCGA-D5-6540 (**l**). **Figure S17** Frequency of MSI events in the 25 states of the chromatin state map averaged across 127 epigenomes for samples TCGA-EW-A1PC (a), TCGA-KL-8330 (b), TCGA-KM-8442 (c), TCGA-EY-A1GS (d), TCGA-KL-8337 (e), TCGA-KM-8443 (f), TCGA-F1-6177 (g), TCGA-KL-8341 (h), TCGA-KO-8409 (i), TCGA-F1-6875 (j), TCGA-KL-8346 (k), TCGA-QG-A5Z2 (l), (**m**) TCGA−IA−A40Y and (**n**) TCGA−KM−8440.

Figure S18 Enrichment of MSI events in the 25 states of the chromatin state map. Odds ratio (**a**) and p values (**b**) for each chromatin state in the 30 whole-genomes displaying the highest MSI rates. Enrichment of MSI events in each chromatin state was tested using Fisher's exact test (Online Methods).

Figure S19 Mutation signatures of predicted MSI-H genomes. The relative frequencies of the 96 trinucleotide mutation contexts (strand symmetric) are shown for MSI-H, MSI-H (predicted) and MSI-L/MSS cases (red, orange and green, respectively). Mutations were pooled across the genomes of three MSI-prone tumor types of COAD/READ, STAD and UCEC. Arrows indicate known mutation features of MSI-H genomes, e.g.*,* C>T transitions in (A/C/G)pCpN sequence contexts and C>A transversions at an CpCpN context, which are enriched in both MSI-H and MSI-H (predicted) genomes compared to MSI-L/MSS genomes.

Supplementary Tables

Table S1 The TCGA barcodes and the total number of MSI calls showing significant differences (FDR < 0.05; Kolmogorov-Smirnov test) in the MS allele lengths between the tumor and matched normal exomes are listed.

Table S2 Results of the DAVID analysis.

Table S3 Frameshift MSI events observed in MSI-H tumors among 13 out the 151 DNA repair genes examined are listed.

Table S4 Frameshift MSI events observed in MSI-H tumors among 12 out of the 130 major cancer-related genes examined are listed.

Table S5 The coding MS loci harboring recurrent frameshift MSI events in COAD, STAD and UCEC MSI-H tumors found using exome sequencing data are listed.

Table S6 Recurrent MSI events in 3'-UTR regions are listed.

Table S7 Recurrent MSI events in 5'-UTR regions are listed.

Table S8 The coding MS repeats harboring recurrent frameshift MSI events in MSS, MSI-L and MSI-H genomes identified using whole exome sequence data from 22 cancer types are listed.

Table S9 Enrichment for frameshit, 3' UTR and 5' UTR MSI events in COAD, UCEC and STAD. The significance was calculated using the one-tailed exact Fisher's text ($\alpha = 0.05$).

Table S10 The barcodes, the total number of significant MSI calls (FDR < 0.05; Kolmogorov-Smirnov test), and the total number of MS loci profiled are listed for the 708 whole-genomes analyzed.

Table S11 List of MS loci discovered in the human mitochondrial DNA.

Table S12 The barcodes corresponding to the samples used in the analysis of MSI in mitochondrial DNA are listed.

Table S13 Top 1,000 most recurrently altered MS loci in COAD, STAD and UCEC MSI-H exomes.

Table S14 Top 1,000 most recurrently altered MS loci in COAD, STAD and UCEC MSI-H whole genomes.

Table S15 Predicted MSI status for 7,096 exomes.

Table S16 The exonic MS loci discovered in the human genome are listed.

Table S17 Genome-wide reference set of MS repeats.

Supplementary File S1 Matrix of covariates used to train the Random Forest models.

Supplementary File S2 Code implemented to predict MSI status using Random Forest models and conformal prediction.

Tables

Table 1. Exome tumors utilized to profile MSI. The Abbreviation column indicates the cancer type abbreviations used throughout the manuscript. The number of cases predicted as MSI-H at a confidence level of 0.75 is indicated with †.

Tumor type	Abbreviation	Samples	MSI-Hs (frequency)
Uterine corpus endometrial carcinoma	UCEC	265	75 (28.3%)
Stomach adenocarcinoma	STAD	292	64 (21.9%)
Colon adenocarcinoma	COAD	271	45 (16.6%)
Rectal adenocarcinoma	READ	76	3; 4† (9.2%)
Adrenal cortical carcinoma	ACC	92	5† (5.4%)
Esophageal carcinoma	ESCA	183	3 (1.6%)
Cervical squamous cell carcinoma	CESC	305	7† (2.3%)
Ovarian cancer	OV	436	14† (3.2%)
Breast cancer	BRCA	922	16† (1.7%)
Cutaneous melanoma	SKCM	109	0† (0.0%)
Head and neck squamous cell carcinoma	HNSC	505	6† (1.2%)
Glioblastoma multiforme	GBM	316	4† (1,3%)
Urothelial bladder cancer	BLCA	368	2† (0.8%)
Liver hepatocellular carcinoma	LIHC	375	11† (2.9%)
Kidney renal clear cell carcinoma	KIRC	377	4† (1.1%)
Lung adenocarcinoma	LUAD	482	1† (0.2%)
Lung squamous cell carcinoma	LUSC	407	5† (1.2%)
Prostate adenocarcinoma	PRAD	497	3† (0.6%)
Pancreatic cancer	PAC	171	2† (1.1%)
Low grade glioma	LGG	514	3† (0.6%)
Papillary kidney carcinoma	KIRP	286	2† (0.7%)
Thyroid cancer	THCA	493	0† (0%)
Pheochromocytoma and paraganglioma	PHCA	176	0† (0%)
Total		**7,919**	**281**

References

1. Aaltonen, L. A. *et al.* Clues to the pathogenesis of familial colorectal cancer. *Science* **260,** 812–6 (1993).
2. Hendriks, Y. M. C. *et al.* Diagnostic approach and management of Lynch syndrome (hereditary nonpolyposis colorectal carcinoma): a guide for clinicians. *CA. Cancer J. Clin.* **56,** 213–25
3. Herman, J. G. *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc. Natl. Acad. Sci.* **95,** 6870–6875 (1998).
4. Ligtenberg, M. J. L. *et al.* Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat. Genet.* **41,** 112–7 (2009).
5. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2257–61 (2006).
6. Jiricny, J. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.* **7,** 335–346 (2006).
7. Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer-the stable evidence. *Nat. Rev. Clin. Oncol.* **7,** 153–62 (2010).
8. Dudley, J. C., Lin, M.-T., Le, D. T. & Eshleman, J. R. Microsatellite Instability as a Biomarker for PD-1 Blockade. *Clin. Cancer Res.* **22,** 813–20 (2016).
9. Gryfe, R. *et al.* Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N. Engl. J. Med.* **342,** 69–77 (2000).
10. Bodmer, W., Bishop, T. & Karran, P. Genetic steps in colorectal cancer. *Nat Genet* **6,** 217–219 (1994).
11. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348,** 124–8 (2015).
12. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* **371,** 2189–2199 (2014).
13. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372,** 2509–2520 (2015).
14. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
15. Kim, T.-M., Laird, P. W. & Park, P. J. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155,** 858–68 (2013).
16. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513,** 202–209 (2014).
17. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
18. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
19. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–26 (2013).
20. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity.

Nat. Commun. **6,** 8971 (2015).

21. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).
22. Kang, J., D'Andrea, A. D. & Kozono, D. A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Natl. Cancer Inst.* **104,** 670–81 (2012).
23. Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* **15,** 181–194 (2015).
24. Hayward, B. E. *et al.* Extensive gene conversion at the *PMS2* DNA mismatch repair locus. *Hum. Mutat.* **28,** 424–430 (2007).
25. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9 (2013).
26. Huang, J. *et al.* APC mutations in colorectal tumors with mismatch repair deficiency. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 9049–54 (1996).
27. Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46,** 1264–1266 (2014).
28. Tougeron, D. *et al.* Tumor-infiltrating lymphocytes in colorectal cancers with microsatellite instability are correlated with the number and spectrum of frameshift mutations. *Mod. Pathol.* **22,** 1186–1195 (2009).
29. Yoshida, Y. *et al.* ALPK2 is crucial for luminal apoptosis and DNA repair-related gene expression in a three-dimensional colonic-crypt model. *Anticancer Res.* **32,** 2301–8 (2012).
30. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47,** 818–821 (2015).
31. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46,** 1160–1165 (2014).
32. Mayr, C. & Bartel, D. P. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138,** 673–684 (2009).
33. 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).
34. Pal, T., Permuth-Wey, J., Kumar, A. & Sellers, T. A. Systematic review and meta-analysis of ovarian cancers: estimation of microsatellite-high frequency and characterization of mismatch repair deficient tumor histology. *Clin. Cancer Res.* **14,** 6847–54 (2008).
35. Altavilla. Microsatellite instability and hMLH1 and hMSH2 expression in renal tumors. *Oncol. Rep.* (2010). doi:10.3892/or_00000938
36. Watson, M. M. C., Berg, M. & Søreide, K. Prevalence and implications of elevated microsatellite alterations at selected tetranucleotides in cancer. *Br. J. Cancer* **111,** 823–7 (2014).
37. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotech* **33,** 364–376 (2015).
38. Larman, T. C. *et al.* Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl. Acad. Sci.* **109,** 14087–14091 (2012).

39. Schwartz, S. & Perucho, M. Somatic mutations in mitochondrial DNA do not associate with nuclear microsatellite instability in gastrointestinal cancer. *Gastroenterology* **119,** 1806–1807 (2000).
40. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
41. Norinder, U., Carlsson, L., Boyer, S. & Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative To Applicability Domain Determination. *J. Chem. Inf. Model.* **54,** 1596–1603 (2014).
42. Murphy, M. A. & Wentzensen, N. Frequency of mismatch repair deficiency in ovarian cancer: a systematic review This article is a US Government work and, as such, is in the public domain of the United States of America. *Int. J. cancer* **129,** 1914–22 (2011).
43. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).
44. Kuismanen, S. A. *et al.* Endometrial and Colorectal Tumors from Patients with Hereditary Nonpolyposis Colon Cancer Display Different Patterns of Microsatellite Instability. *Am. J. Pathol.* **160,** 1953–1958 (2002).
45. Miller, S. L., Antico, G., Raghunath, P. N., Tomaszewski, J. E. & Clevenger, C. V. Nek3 kinase regulates prolactin-mediated cytoskeletal reorganization and motility of breast cancer cells. *Oncogene* **26,** 4668–4678 (2007).
46. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30,** 1015–6 (2014).
47. Huang, M. N. *et al.* MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci. Rep.* **5,** 13321 (2015).
48. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36 ,** e105–e105 (2008).
49. Bacher, J. W. *et al.* Development of a Fluorescent Multiplex Assay for Detection of MSI-High Tumors. *Dis. Markers* **20,** 237–250 (2004).
50. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–9 (2013).
51. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
52. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
53. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
54. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
55. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer New York, 2013). doi:10.1007/978-1-4614-6849-3

Figure 1

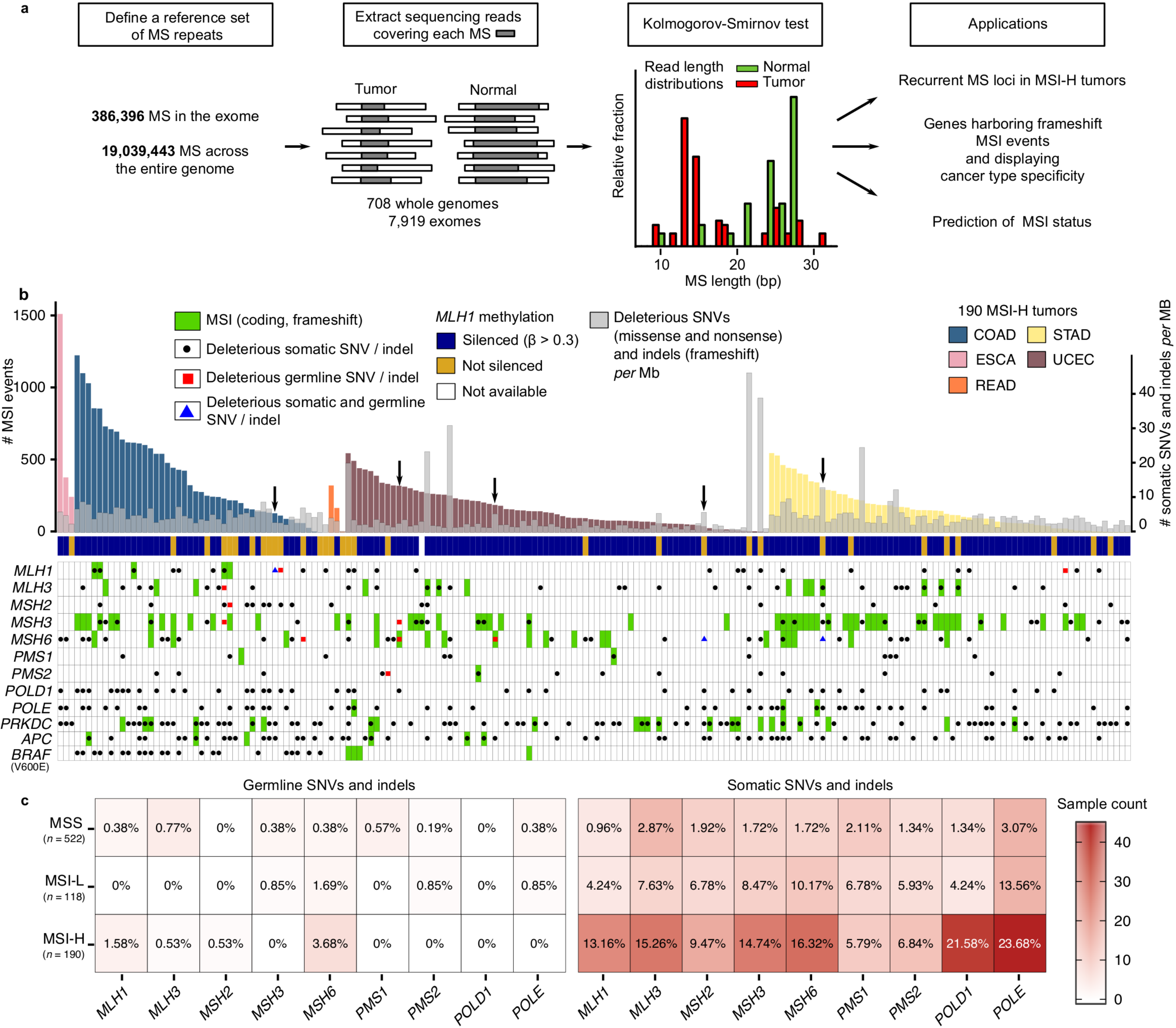

a
Define a reference set of MS repeats
Extract sequencing reads covering each MS
Kolmogorov-Smirnov test
Applications
386,396 MS in the exome
19,039,443 MS across the entire genome
Tumor
Normal
708 whole genomes
7,919 exomes
Read length distributions
Normal
Tumor
Relative fraction
10
20
30
MS length (bp)
Recurrent MS loci in MSI-H tumors
Genes harboring frameshift MSI events and displaying cancer type specificity
Prediction of MSI status
b
# MSI events
1500
1000
500
0
MSI (coding, frameshift)
Deleterious somatic SNV / indel
Deleterious germline SNV / indel
Deleterious somatic and germline SNV / indel
MLH1 methylation
Silenced (β > 0.3)
Not silenced
Not available
Deleterious SNVs (missense and nonsense) and indels (frameshift) per Mb
190 MSI-H tumors
COAD
STAD
ESCA
UCEC
READ
# somatic SNVs and indels per MB
40
30
20
10
0
MLH1
MLH3
MSH2
MSH3
MSH6
PMS1
PMS2
POLD1
POLE
PRKDC
APC
BRAF (V600E)
c
Germline SNVs and indels
Somatic SNVs and indels
Sample count
40
30
20
10
0
MSS (n = 522)
MSI-L (n = 118)
MSI-H (n = 190)
MLH1
MLH3
MSH2
MSH3
MSH6
PMS1
PMS2
POLD1
POLE
MSS germline: 0.38% 0.77% 0% 0.38% 0.38% 0.57% 0.19% 0% 0.38%
MSS somatic: 0.96% 2.87% 1.92% 1.72% 1.72% 2.11% 1.34% 1.34% 3.07%
MSI-L germline: 0% 0% 0% 0.85% 1.69% 0% 0.85% 0% 0.85%
MSI-L somatic: 4.24% 7.63% 6.78% 8.47% 10.17% 6.78% 5.93% 4.24% 13.56%
MSI-H germline: 1.58% 0.53% 0.53% 0% 3.68% 0% 0% 0% 0%
MSI-H somatic: 13.16% 15.26% 9.47% 14.74% 16.32% 5.79% 6.84% 21.58% 23.68%

Figure 2

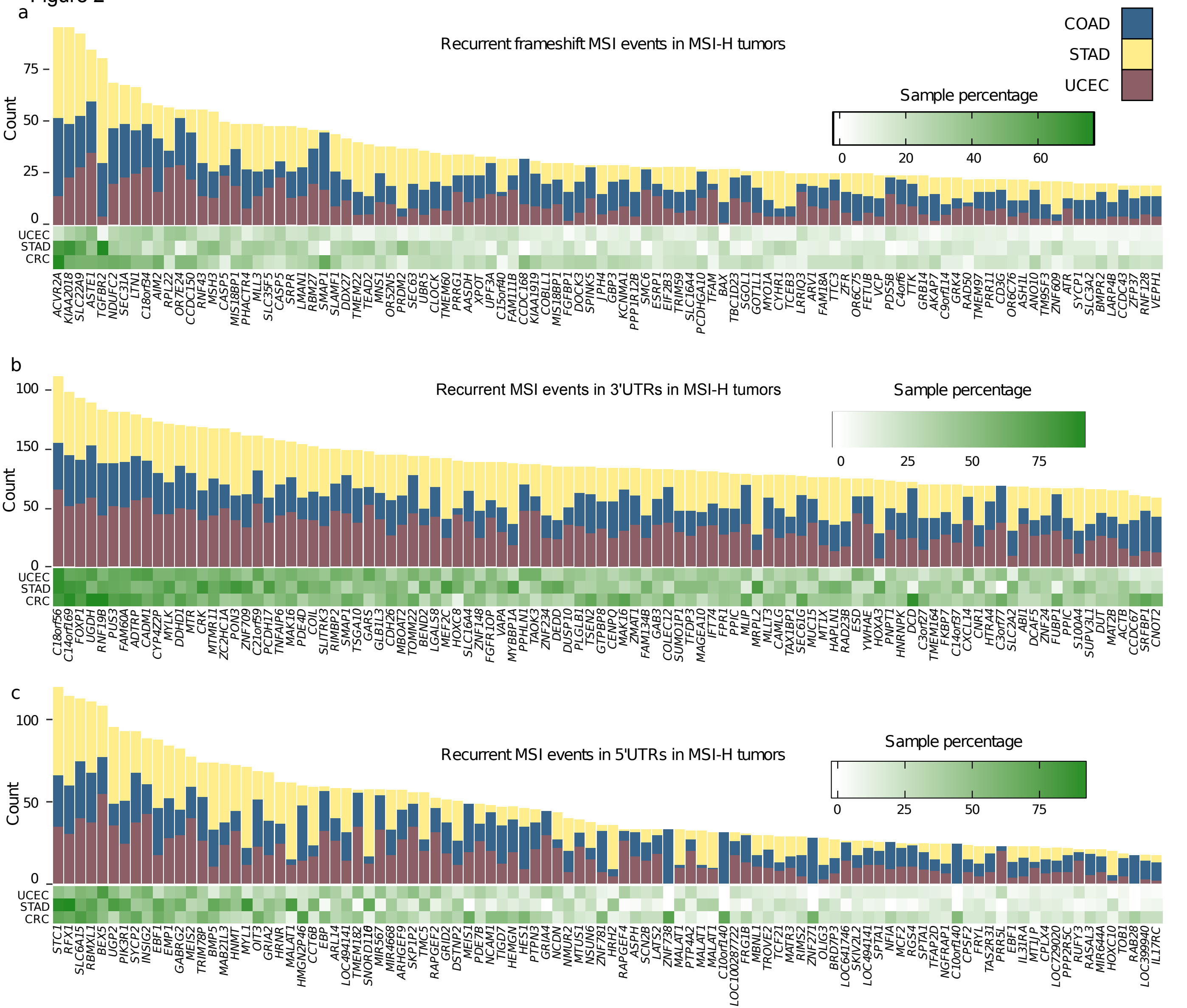

a
Recurrent frameshift MSI events in MSI-H tumors
Count
0
25
50
75
COAD
STAD
UCEC
Sample percentage
0
20
40
60
UCEC
STAD
CRC
ACVR2A KIAA2018 SLC22A9 ASTE1 TGFBR2 NDUFC2 SEC31A LTN1 C18orf34 AIM2 RPL22 OR7E24 CCDC150 RNF43 MSH3 CASP5 MIS18BP1 PHACTR4 MLL3 SLC35F5 CASP5 SRPR LMAN1 RBM27 SMAP1 SLAMF1 DDX27 TMEM22 TEAD2 MNS1 OR52N5 PRDM2 SEC63 UBR5 CLOCK TMEM60 PRRG1 AASDH XPOT UPF3A C15orf40 FAM111B CCDC168 KIAA1919 COBLL1 MIS18BP1 FGFBP1 DOCK3 SPINK5 JPH4 GBP3 KCNMA1 PPP1R12B SMC6 ESRP1 EIF2B3 TRIM59 SLC16A4 PCDHGA10 TFAM BAX TBC1D23 SGOL1 GOT1L1 MYO1A CYHR1 TCEB3 LRRIQ3 ARV1 FAM18A TTC3 ZFR OR6C75 FETUB VCP PDS5B C4orf6 TTK GRB14 AKAP7 C9orf114 GRK4 RAD50 TMEM97 PRR11 CD3G OR6C76 ASH1L ANO10 TM9SF3 ZNF609 ATR SYCP1 SLC3A2 BMPR2 LARP4B CCDC43 ZFP37 RNF128 VEPH1
b
Recurrent MSI events in 3'UTRs in MSI-H tumors
Count
0
50
150
100
Sample percentage
0
25
50
75
UCEC
STAD
CRC
C18orf56 C14orf169 FOXP1 UGDH RNF19B PUS3 FAM60A ADTRP CADM1 CYP4Z2P MYLK DDHD1 MTR CRK MTMR11 ZC2HC1A PON3 ZNF709 C21orf59 PCDH17 TNFAIP6 MAK16 PDE4D COIL SLITRK3 RIMBP2 SMAP1 TSGA10 GARS GLB1L3 CDH26 MBOAT2 TOMM22 BEND2 LGALS8 MEF2C HOXC8 SLC16A4 ZNF148 FGFR1OP VAPA MYBBP1A PPHLN1 TAOK3 ZNF234 DEDD DUSP10 PLGLB1 TSEN2 GTPBP8 CENPQ MAK16 ZMAT1 FAM134B GAB3 COLEC12 SUMO1P1 TFDP3 MAGEA10 IFT74 FPR1 PPIC MLIP MRPL2 MLLT3 CAMLG TAX1BP1 SEC61G MUC15 MT1X HAPLN1 RAD23B ESD YWHAE HOXA3 PNPT1 HNRNPK DCD C3orf27 TMEM164 FKBP7 C14orf37 CXCL14 CNR1 HTRA4 C3orf77 SLC2A2 ABI1 DCAF5 ZNF24 FUBP1 PPIC S100A4 SUPV3L1 DUT MAT2B ACTB CCDC67 SRFBP1 CNOT2
c
Recurrent MSI events in 5'UTRs in MSI-H tumors
Count
0
50
100
Sample percentage
0
25
50
75
UCEC
STAD
CRC
STC1 RFX1 SLC6A15 RBMXL1 BEX5 UGP2 PIK3R1 SYCP2 INSIG2 EBF1 EMP1 GABRG2 MEIS2 TRIM78P BMP5 MAB21L3 HNMT MYL1 OIT3 GRIA2 HRNR MALAT1 HMGN2P46 CCT6B EBP ARL14 LOC494141 TMEM182 SNORD116 MIR567 MIR4668 ARHGEF9 SKP1P2 TMC5 RAPGEF2 GRID2 DSTNP2 MEIS1 PDE7B NCAM1 TIGD7 HEMGN HES1 PTPN3 GRIA4 NCDN NMUR2 MTUS1 NSUN6 ZNF781 HRH2 RAPGEF4 ASPH SCN2B LATS2 ZNF738 MALAT1 PTP4A2 MALAT1 MALAT1 C10orf140 LOC100287722 FRG1B MBNL1 TROVE2 TCF21 MATR3 RIMS2 ZNF271 OLIG3 BRD7P3 LOC641746 SKIV2L2 LOC494141 SPTA1 NFIA MCF2 RGS4 SPTA1 TFAP2D NGFRAP1 C10orf140 CPSF2 FRYL TAS2R31 PRR5L EBF1 IL31RA MT1JP CPLX4 LOC729020 PPP2R5C RUFY3 RASAL3 MIR644A HOXC10 TOB1 RAB28 LOC399040 IL17RC

Figure 3

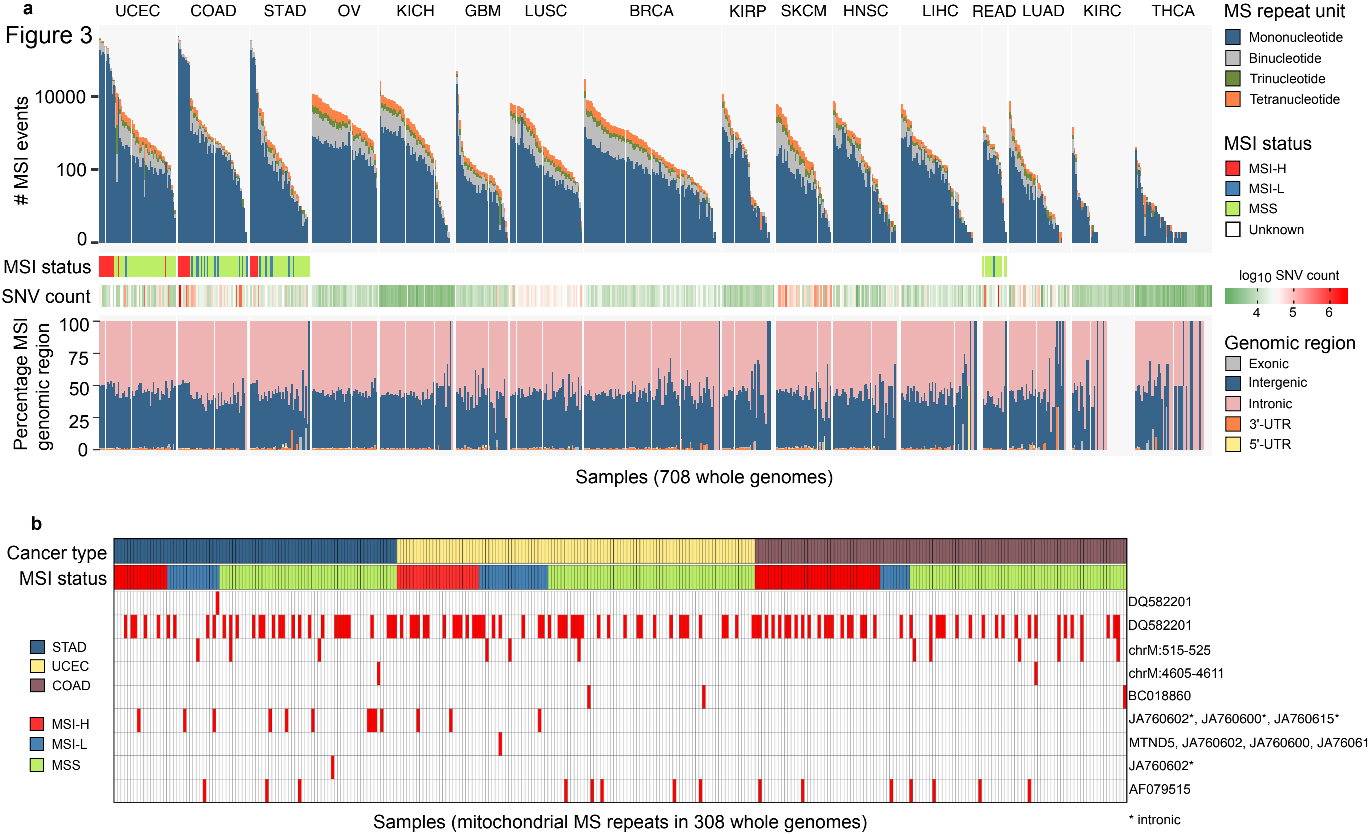

a
UCEC
COAD
STAD
OV
KICH
GBM
LUSC
BRCA
KIRP
SKCM
HNSC
LIHC
READ
LUAD
KIRC
THCA
# MSI events
10000
100
0
MSI status
SNV count
Percentage MSI genomic region
100
75
50
25
0
Samples (708 whole genomes)
MS repeat unit
Mononucleotide
Binucleotide
Trinucleotide
Tetranucleotide
MSI status
MSI-H
MSI-L
MSS
Unknown
$\log_{10}$ SNV count
4
5
6
Genomic region
Exonic
Intergenic
Intronic
3'-UTR
5'-UTR
b
Cancer type
MSI status
STAD
UCEC
COAD
MSI-H
MSI-L
MSS
DQ582201
DQ582201
chrM:515-525
chrM:4605-4611
BC018860
JA760602*, JA760600*, JA760615*
MTND5, JA760602, JA760600, JA76061
JA760602*
AF079515
Samples (mitochondrial MS repeats in 308 whole genomes)
* intronic

Figure 4

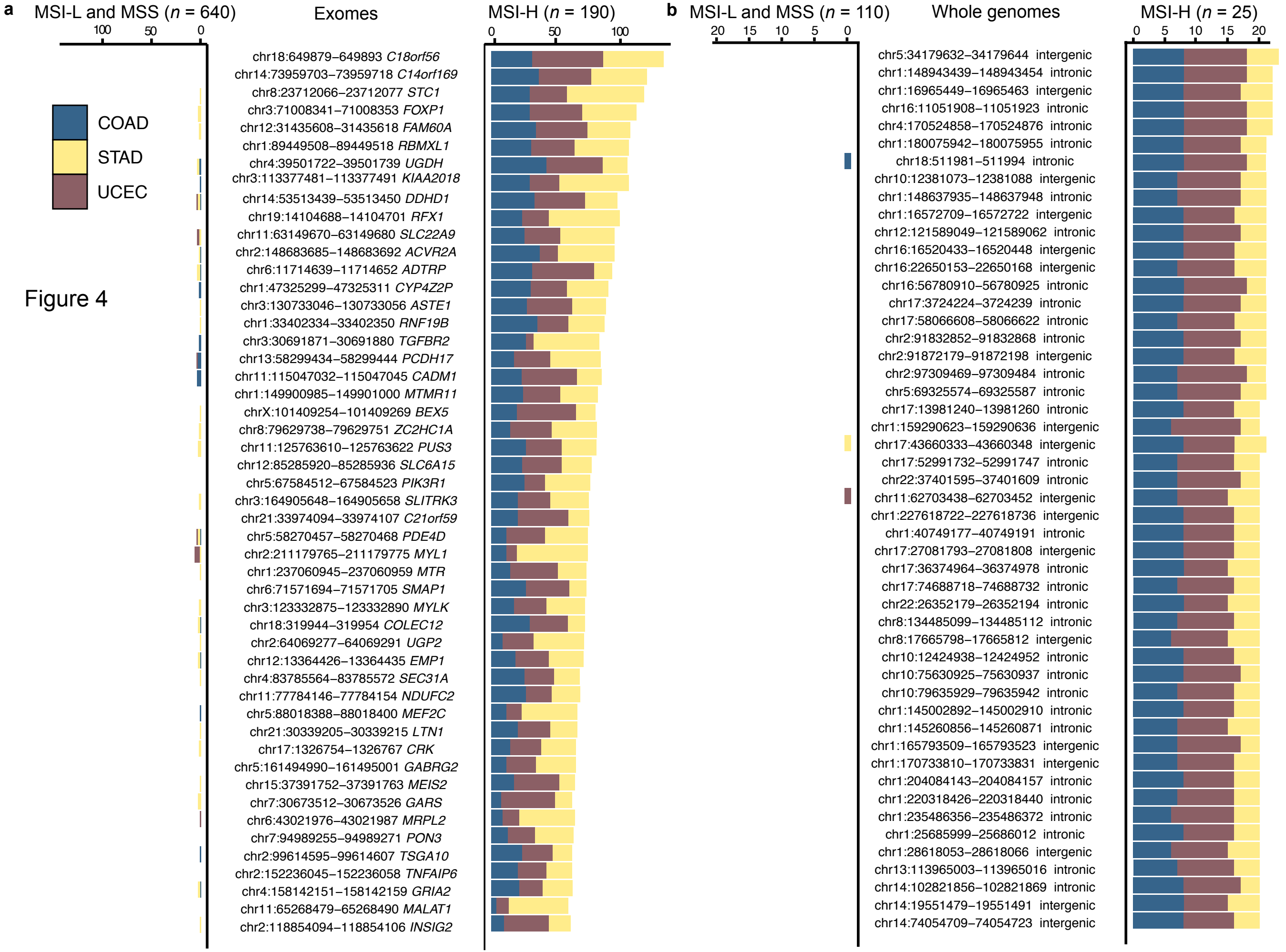
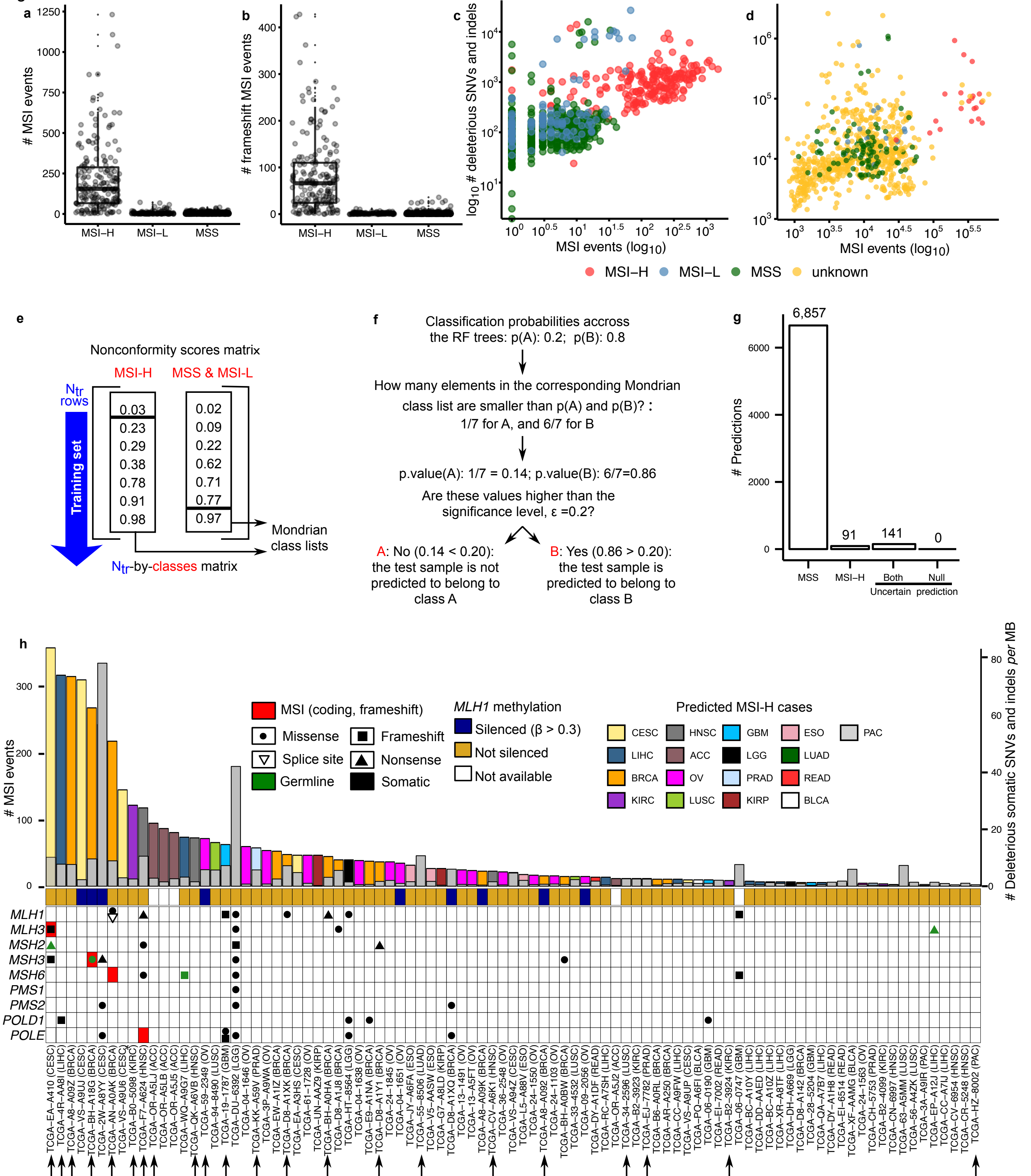
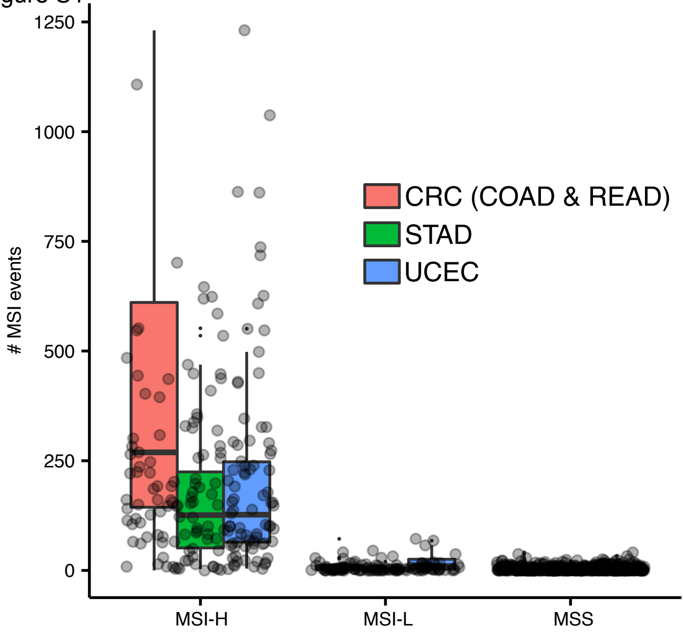

a
MSI-L and MSS (n = 640)
100
50
0
Exomes
MSI-H (n = 190)
0
50
100
COAD
STAD
UCEC
chr18:649879–649893 C18orf56
chr14:73959703–73959718 C14orf169
chr8:23712066–23712077 STC1
chr3:71008341–71008353 FOXP1
chr12:31435608–31435618 FAM60A
chr1:89449508–89449518 RBMXL1
chr4:39501722–39501739 UGDH
chr3:113377481–113377491 KIAA2018
chr14:53513439–53513450 DDHD1
chr19:14104688–14104701 RFX1
chr11:63149670–63149680 SLC22A9
chr2:148683685–148683692 ACVR2A
chr6:11714639–11714652 ADTRP
chr1:47325299–47325311 CYP4Z2P
chr3:130733046–130733056 ASTE1
chr1:33402334–33402350 RNF19B
chr3:30691871–30691880 TGFBR2
chr13:58299434–58299444 PCDH17
chr11:115047032–115047045 CADM1
chr1:149900985–149901000 MTMR11
chrX:101409254–101409269 BEX5
chr8:79629738–79629751 ZC2HC1A
chr11:125763610–125763622 PUS3
chr12:85285920–85285936 SLC6A15
chr5:67584512–67584523 PIK3R1
chr3:164905648–164905658 SLITRK3
chr21:33974094–33974107 C21orf59
chr5:58270457–58270468 PDE4D
chr2:211179765–211179775 MYL1
chr1:237060945–237060959 MTR
chr6:71571694–71571705 SMAP1
chr3:123332875–123332890 MYLK
chr18:319944–319954 COLEC12
chr2:64069277–64069291 UGP2
chr12:13364426–13364435 EMP1
chr4:83785564–83785572 SEC31A
chr11:77784146–77784154 NDUFC2
chr5:88018388–88018400 MEF2C
chr21:30339205–30339215 LTN1
chr17:1326754–1326767 CRK
chr5:161494990–161495001 GABRG2
chr15:37391752–37391763 MEIS2
chr7:30673512–30673526 GARS
chr6:43021976–43021987 MRPL2
chr7:94989255–94989271 PON3
chr2:99614595–99614607 TSGA10
chr2:152236045–152236058 TNFAIP6
chr4:158142151–158142159 GRIA2
chr11:65268479–65268490 MALAT1
chr2:118854094–118854106 INSIG2
b
MSI-L and MSS (n = 110)
20
15
10
5
0
Whole genomes
MSI-H (n = 25)
0
5
10
15
20
chr5:34179632–34179644 intergenic
chr1:148943439–148943454 intronic
chr1:16965449–16965463 intergenic
chr16:11051908–11051923 intronic
chr4:170524858–170524876 intronic
chr1:180075942–180075955 intronic
chr18:511981–511994 intronic
chr10:12381073–12381088 intergenic
chr1:148637935–148637948 intronic
chr1:16572709–16572722 intergenic
chr12:121589049–121589062 intronic
chr16:16520433–16520448 intergenic
chr16:22650153–22650168 intergenic
chr16:56780910–56780925 intronic
chr17:3724224–3724239 intronic
chr17:58066608–58066622 intronic
chr2:91832852–91832868 intronic
chr2:91872179–91872198 intergenic
chr2:97309469–97309484 intronic
chr5:69325574–69325587 intronic
chr17:13981240–13981260 intronic
chr1:159290623–159290636 intergenic
chr17:43660333–43660348 intergenic
chr17:52991732–52991747 intronic
chr22:37401595–37401609 intronic
chr11:62703438–62703452 intergenic
chr1:227618722–227618736 intergenic
chr1:40749177–40749191 intronic
chr17:27081793–27081808 intergenic
chr17:36374964–36374978 intronic
chr17:74688718–74688732 intronic
chr22:26352179–26352194 intronic
chr8:134485099–134485112 intronic
chr8:17665798–17665812 intergenic
chr10:12424938–12424952 intronic
chr10:75630925–75630937 intronic
chr10:79635929–79635942 intronic
chr1:145002892–145002910 intronic
chr1:145260856–145260871 intronic
chr1:165793509–165793523 intergenic
chr1:170733810–170733831 intergenic
chr1:204084143–204084157 intronic
chr1:220318426–220318440 intronic
chr1:235486356–235486372 intronic
chr1:25685999–25686012 intronic
chr1:28618053–28618066 intergenic
chr13:113965003–113965016 intronic
chr14:102821856–102821869 intronic
chr14:19551479–19551491 intergenic
chr14:74054709–74054723 intergenic

Figure 5



a
# MSI events
1250
1000
750
500
250
0
MSI–H
MSI–L
MSS
b
# frameshift MSI events
400
300
200
100
0
MSI–H
MSI–L
MSS
c
Exomes
log10 # deleterious SNVs and indels
10^1
10^2
10^3
10^4
10^0
$10^{0.5}$
10^1
$10^{1.5}$
10^2
$10^{2.5}$
10^3
MSI events (log10)
d
Whole genomes
10^3
10^4
10^5
10^6
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
$10^{5.5}$
MSI events (log10)
MSI–H
MSI–L
MSS
unknown
e
Nonconformity scores matrix
MSI-H
MSS & MSI-L
Ntr rows
0.03
0.23
0.29
0.38
0.78
0.91
0.98
0.02
0.09
0.22
0.62
0.71
0.77
0.97
Training set
Mondrian class lists
Ntr-by-classes matrix
f
Classification probabilities accross the RF trees: p(A): 0.2; p(B): 0.8
How many elements in the corresponding Mondrian class list are smaller than p(A) and p(B)? : 1/7 for A, and 6/7 for B
p.value(A): 1/7 = 0.14; p.value(B): 6/7=0.86
Are these values higher than the significance level, ε =0.2?
A: No (0.14 < 0.20): the test sample is not predicted to belong to class A
B: Yes (0.86 > 0.20): the test sample is predicted to belong to class B
g
# Predictions
6000
4000
2000
0
6,857
91
141
0
MSS
MSI–H
Both
Null
Uncertain prediction
h
# MSI events
300
200
100
0
# Deleterious somatic SNVs and indels per MB
80
60
40
20
0
MSI (coding, frameshift)
Missense
Frameshift
Splice site
Nonsense
Germline
Somatic
MLH1 methylation
Silenced (β > 0.3)
Not silenced
Not available
Predicted MSI-H cases
CESC
HNSC
GBM
ESO
PAC
LIHC
ACC
LGG
LUAD
BRCA
OV
PRAD
READ
KIRC
LUSC
KIRP
BLCA
MLH1
MLH3
MSH2
MSH3
MSH6
PMS1
PMS2
POLD1
POLE
TCGA–EA–A410 (CESC)
TCGA–4R–AA8I (LIHC)
TCGA–A8–A09Z (BRCA)
TCGA–VS–A9UJ (CESC)
TCGA–BH–A18G (BRCA)
TCGA–2W–A8YY (CESC)
TCGA–AN–A0AK (BRCA)
TCGA–VS–A9U6 (CESC)
TCGA–B0–5098 (KIRC)
TCGA–F7–A624 (HNSC)
TCGA–OR–A5LJ (ACC)
TCGA–OR–A5LB (ACC)
TCGA–OR–A5J5 (ACC)
TCGA–WQ–A9G7 (LIHC)
TCGA–QK–A6VB (HNSC)
TCGA–59–2349 (OV)
TCGA–94–8490 (LUSC)
TCGA–19–1787 (GBM)
TCGA–DU–6392 (LGG)
TCGA–04–1646 (OV)
TCGA–KK–A59V (PRAD)
TCGA–3P–A9WA (OV)
TCGA–EW–A1IZ (BRCA)
TCGA–D8–A1XK (BRCA)
TCGA–EA–A3HS (CESC)
TCGA–61–1728 (OV)
TCGA–UN–AAZ9 (KIRP)
TCGA–BH–A0HA (BRCA)
TCGA–D8–A1J8 (BRCA)
TCGA–HT–8564 (LGG)
TCGA–04–1638 (OV)
TCGA–E9–A1NA (BRCA)
TCGA–D8–A1Y1 (BRCA)
TCGA–24–1845 (OV)
TCGA–04–1651 (OV)
TCGA–JY–A6FA (ESO)
TCGA–55–8506 (LUAD)
TCGA–V5–AASW (ESO)
TCGA–G7–A8LD (KIRP)
TCGA–D8–A1XQ (BRCA)
TCGA–13–1491 (OV)
TCGA–13–A5FT (OV)
TCGA–A8–A09K (BRCA)
TCGA–CV–A6K1 (HNSC)
TCGA–36–2548 (OV)
TCGA–VS–A94Z (CESC)
TCGA–L5–A88V (ESO)
TCGA–24–1550 (OV)
TCGA–A8–A092 (BRCA)
TCGA–24–1103 (OV)
TCGA–BH–A0BW (BRCA)
TCGA–33–4532 (LUSC)
TCGA–09–2056 (OV)
TCGA–DY–A1DF (READ)
TCGA–RC–A7SH (LIHC)
TCGA–OR–A5J2 (ACC)
TCGA–34–2596 (LUSC)
TCGA–B2–3923 (KIRC)
TCGA–EJ–7782 (PRAD)
TCGA–B6–A0RL (BRCA)
TCGA–AR–A250 (BRCA)
TCGA–CC–A9FW (LIHC)
TCGA–VS–A9UP (CESC)
TCGA–PQ–A6FI (BLCA)
TCGA–06–0190 (GBM)
TCGA–EI–7002 (READ)
TCGA–B2–3924 (KIRC)
TCGA–06–0747 (GBM)
TCGA–BC–A10Y (LIHC)
TCGA–DD–AAED (LIHC)
TCGA–BC–A10Z (LIHC)
TCGA–XR–A8TF (LIHC)
TCGA–DH–A669 (LGG)
TCGA–D8–A142 (BRCA)
TCGA–28–5204 (GBM)
TCGA–QA–A7B7 (LIHC)
TCGA–DY–A1H8 (READ)
TCGA–EI–6514 (READ)
TCGA–XF–AAMG (BLCA)
TCGA–24–1563 (OV)
TCGA–CH–5753 (PRAD)
TCGA–B2–4099 (KIRC)
TCGA–CN–6997 (HNSC)
TCGA–63–A5MM (LUSC)
TCGA–56–A4ZK (LUSC)
TCGA–3A–A9IR (PAC)
TCGA–EP–A12J (LIHC)
TCGA–CC–A7IJ (LIHC)
TCGA–CV–6954 (HNSC)
TCGA–CR–5248 (HNSC)
TCGA–HZ–8002 (PAC)

Figure S1



# MSI events
1250
1000
750
500
250
0
MSI-H
MSI-L
MSS
🟥 CRC (COAD & READ)
🟩 STAD
🟦 UCEC

Figure S2

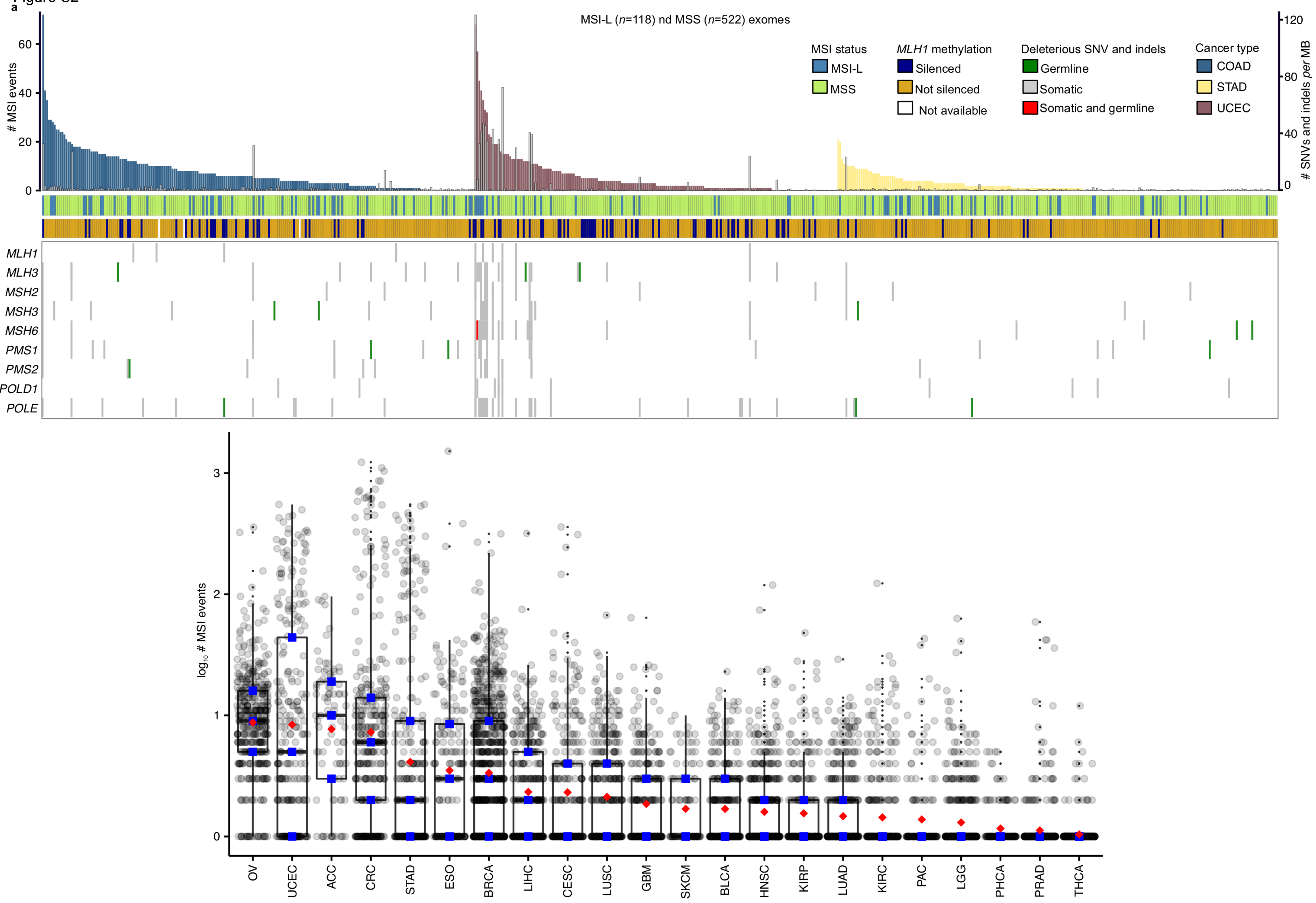

a
MSI-L (n=118) nd MSS (n=522) exomes
# MSI events
0
20
40
60
# SNVs and indels per MB
0
40
80
120
MSI status
MSI-L
MSS
MLH1 methylation
Silenced
Not silenced
Not available
Deleterious SNV and indels
Germline
Somatic
Somatic and germline
Cancer type
COAD
STAD
UCEC
MLH1
MLH3
MSH2
MSH3
MSH6
PMS1
PMS2
POLD1
POLE
log10 # MSI events
0
1
2
3
OV
UCEC
ACC
CRC
STAD
ESO
BRCA
LIHC
CESC
LUSC
GBM
SKCM
BLCA
HNSC
KIRP
LUAD
KIRC
PAC
LGG
PHCA
PRAD
THCA

Figure S3

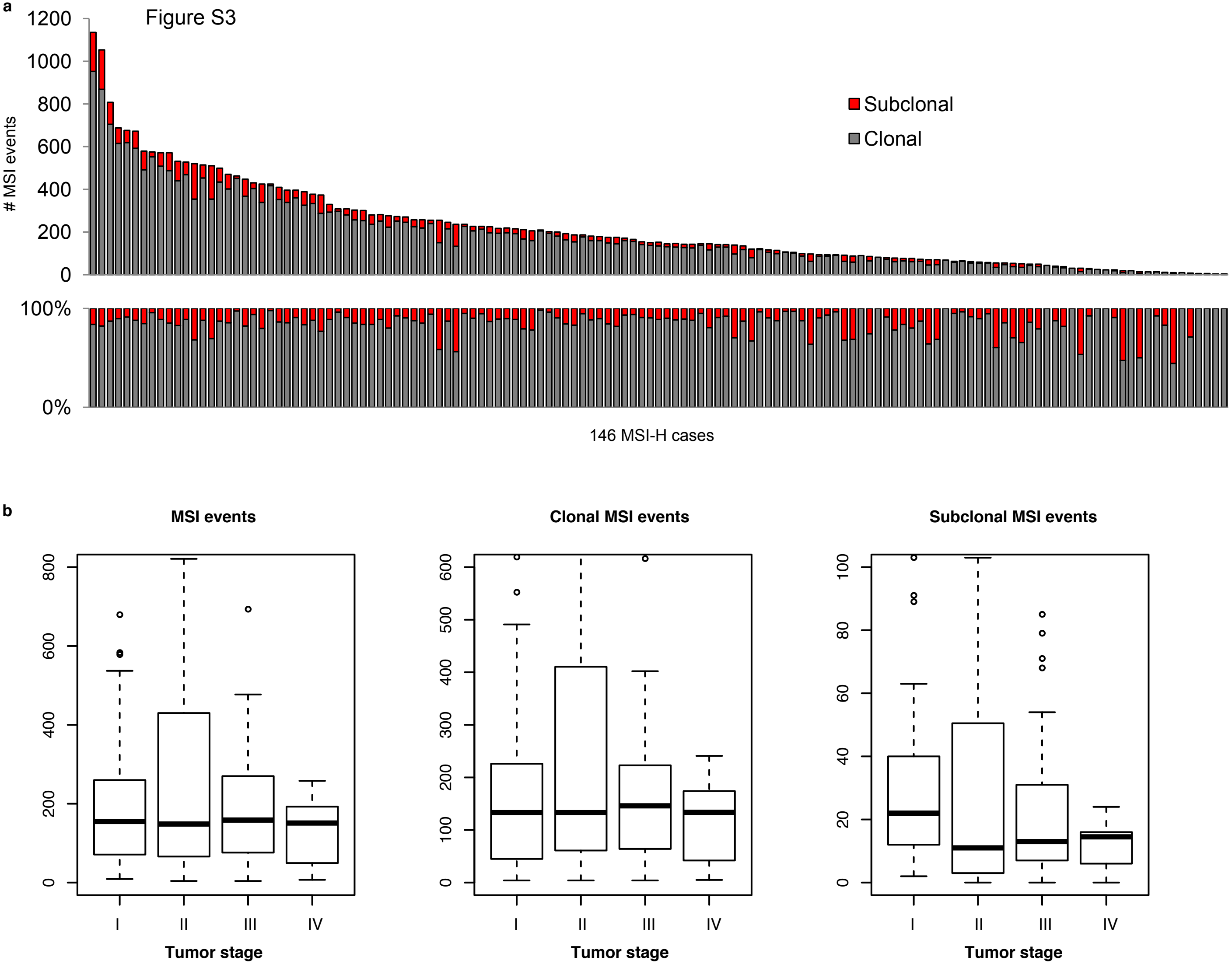

a
1200
1000
800
600
400
200
0
# MSI events
🟥 Subclonal
▨ Clonal
100%
0%
146 MSI-H cases
b
MSI events
800
600
400
200
0
I
II
III
IV
Tumor stage
Clonal MSI events
600
500
400
300
200
100
0
I
II
III
IV
Tumor stage
Subclonal MSI events
100
80
60
40
20
0
I
II
III
IV
Tumor stage

Figure S4

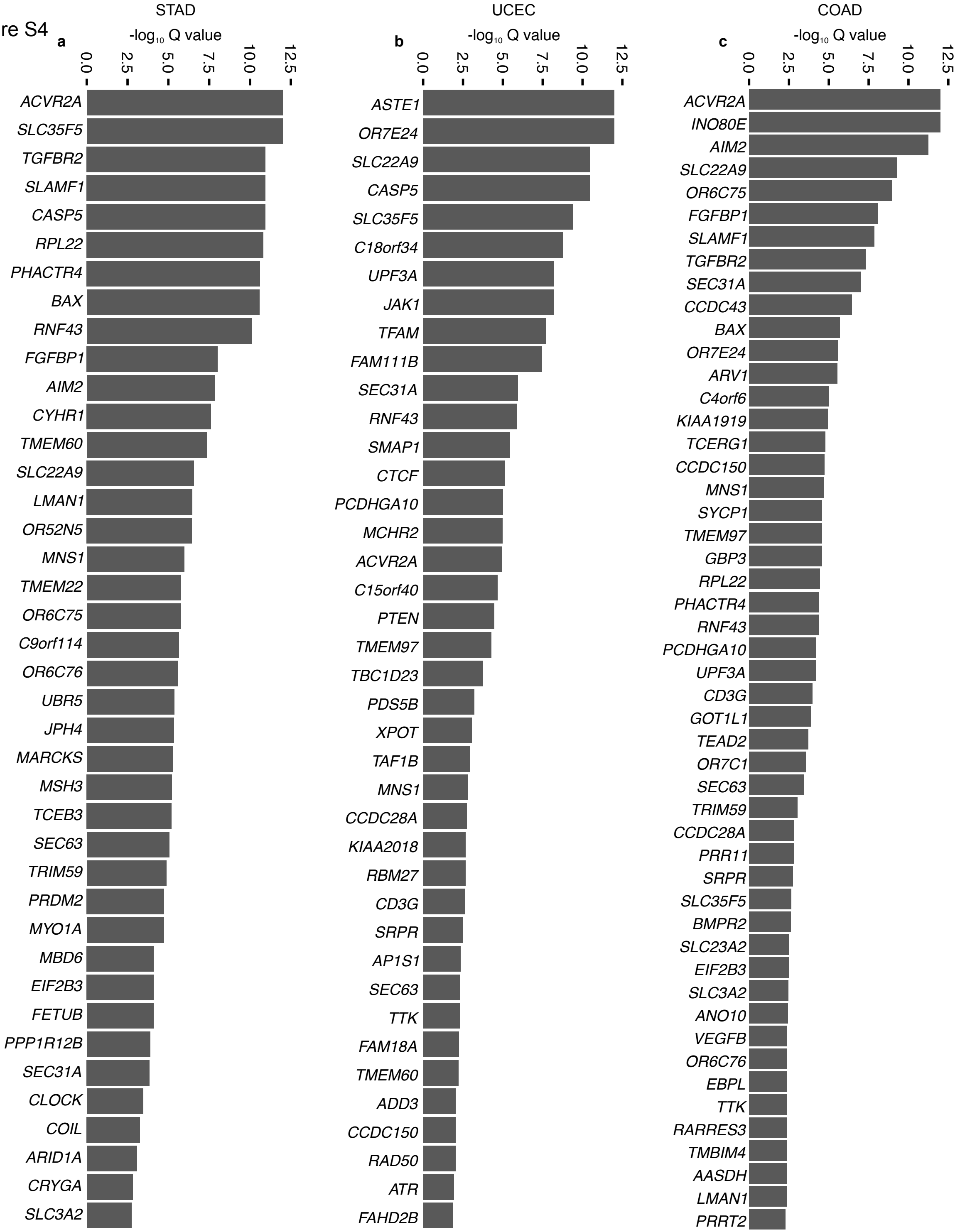

a
STAD
-log10 Q value
0.0
2.5
5.0
7.5
10.0
12.5
ACVR2A
SLC35F5
TGFBR2
SLAMF1
CASP5
RPL22
PHACTR4
BAX
RNF43
FGFBP1
AIM2
CYHR1
TMEM60
SLC22A9
LMAN1
OR52N5
MNS1
TMEM22
OR6C75
C9orf114
OR6C76
UBR5
JPH4
MARCKS
MSH3
TCEB3
SEC63
TRIM59
PRDM2
MYO1A
MBD6
EIF2B3
FETUB
PPP1R12B
SEC31A
CLOCK
COIL
ARID1A
CRYGA
SLC3A2
b
UCEC
-log10 Q value
0.0
2.5
5.0
7.5
10.0
12.5
ASTE1
OR7E24
SLC22A9
CASP5
SLC35F5
C18orf34
UPF3A
JAK1
TFAM
FAM111B
SEC31A
RNF43
SMAP1
CTCF
PCDHGA10
MCHR2
ACVR2A
C15orf40
PTEN
TMEM97
TBC1D23
PDS5B
XPOT
TAF1B
MNS1
CCDC28A
KIAA2018
RBM27
CD3G
SRPR
AP1S1
SEC63
TTK
FAM18A
TMEM60
ADD3
CCDC150
RAD50
ATR
FAHD2B
c
COAD
-log10 Q value
0.0
2.5
5.0
7.5
10.0
12.5
ACVR2A
INO80E
AIM2
SLC22A9
OR6C75
FGFBP1
SLAMF1
TGFBR2
SEC31A
CCDC43
BAX
OR7E24
ARV1
C4orf6
KIAA1919
TCERG1
CCDC150
MNS1
SYCP1
TMEM97
GBP3
RPL22
PHACTR4
RNF43
PCDHGA10
UPF3A
CD3G
GOT1L1
TEAD2
OR7C1
SEC63
TRIM59
CCDC28A
PRR11
SRPR
SLC35F5
BMPR2
SLC23A2
EIF2B3
SLC3A2
ANO10
VEGFB
OR6C76
EBPL
TTK
RARRES3
TMBIM4
AASDH
LMAN1
PRRT2

Figure S5

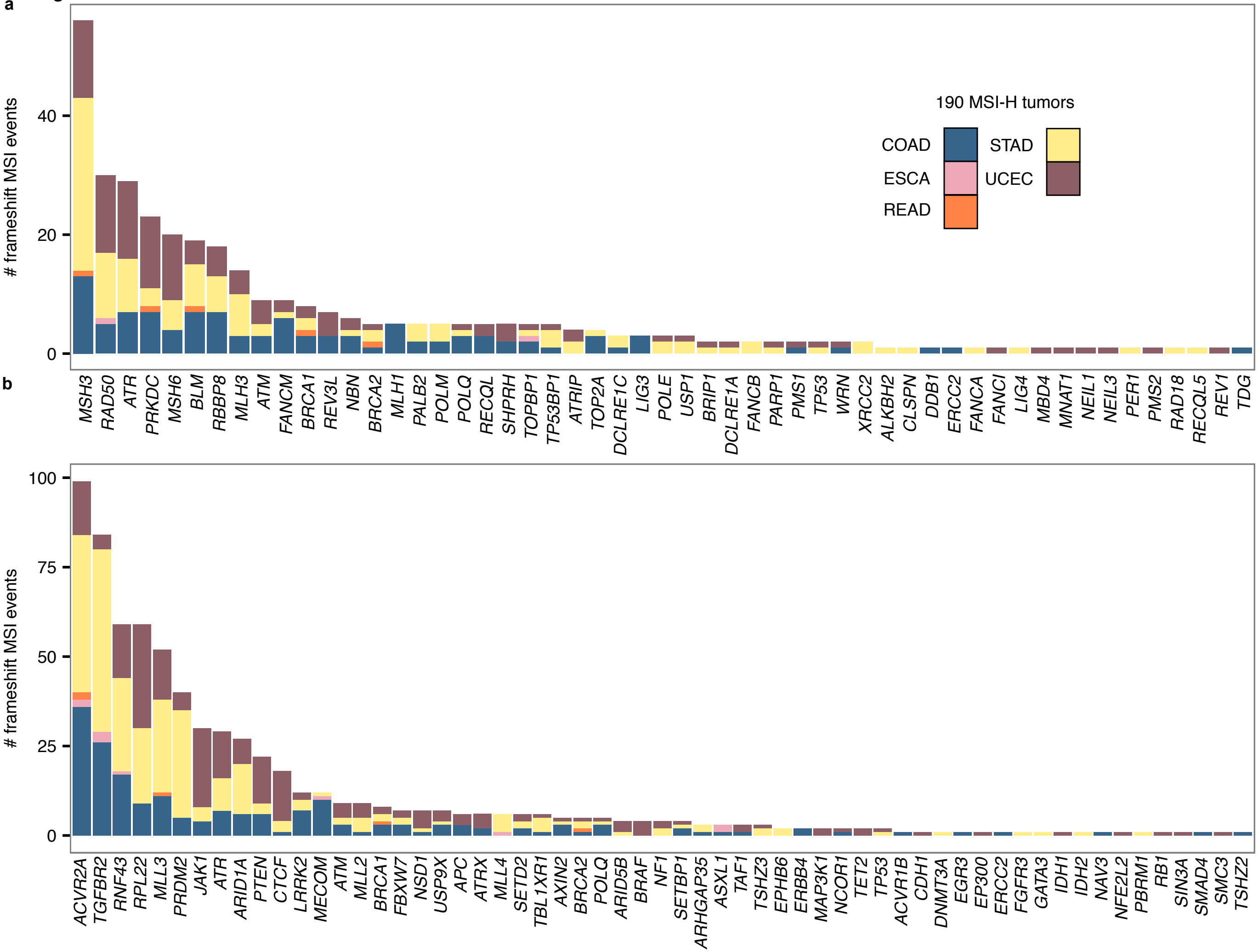

a
# frameshift MSI events
0
20
40
190 MSI-H tumors
COAD
ESCA
READ
STAD
UCEC
MSH3
RAD50
ATR
PRKDC
MSH6
BLM
RBBP8
MLH3
ATM
FANCM
BRCA1
REV3L
NBN
BRCA2
MLH1
PALB2
POLM
POLQ
RECQL
SHPRH
TOPBP1
TP53BP1
ATRIP
TOP2A
DCLRE1C
LIG3
POLE
USP1
BRIP1
DCLRE1A
FANCB
PARP1
PMS1
TP53
WRN
XRCC2
ALKBH2
CLSPN
DDB1
ERCC2
FANCA
FANCI
LIG4
MBD4
MNAT1
NEIL1
NEIL3
PER1
PMS2
RAD18
RECQL5
REV1
TDG
b
# frameshift MSI events
0
25
50
75
100
ACVR2A
TGFBR2
RNF43
RPL22
MLL3
PRDM2
JAK1
ATR
ARID1A
PTEN
CTCF
LRRK2
MECOM
ATM
MLL2
BRCA1
FBXW7
NSD1
USP9X
APC
ATRX
MLL4
SETD2
TBL1XR1
AXIN2
BRCA2
POLQ
ARID5B
BRAF
NF1
SETBP1
ARHGAP35
ASXL1
TAF1
TSHZ3
EPHB6
ERBB4
MAP3K1
NCOR1
TET2
TP53
ACVR1B
CDH1
DNMT3A
EGR3
EP300
ERCC2
FGFR3
GATA3
IDH1
IDH2
NAV3
NFE2L2
PBRM1
RB1
SIN3A
SMAD4
SMC3
TSHZ2

Figure S6

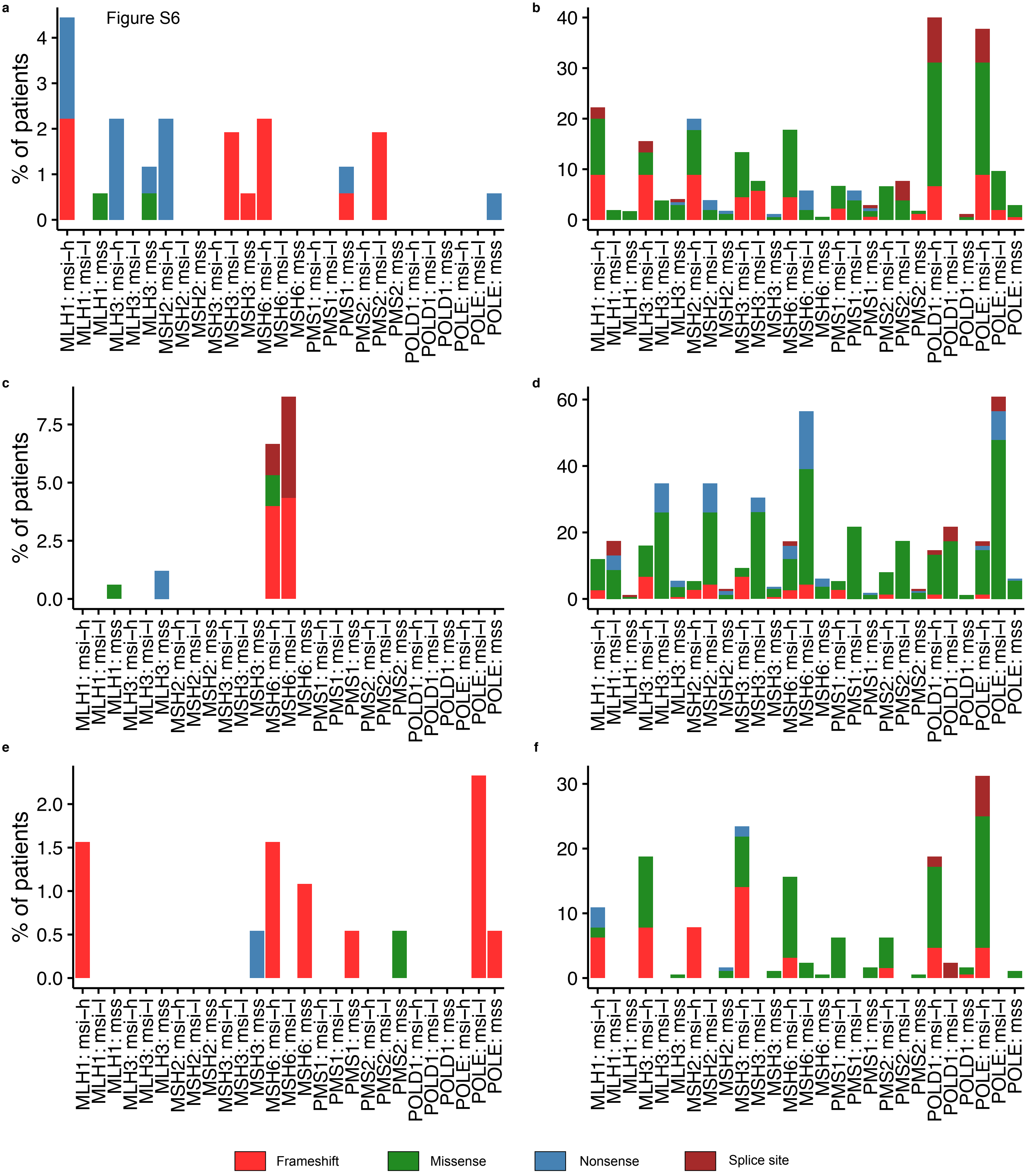
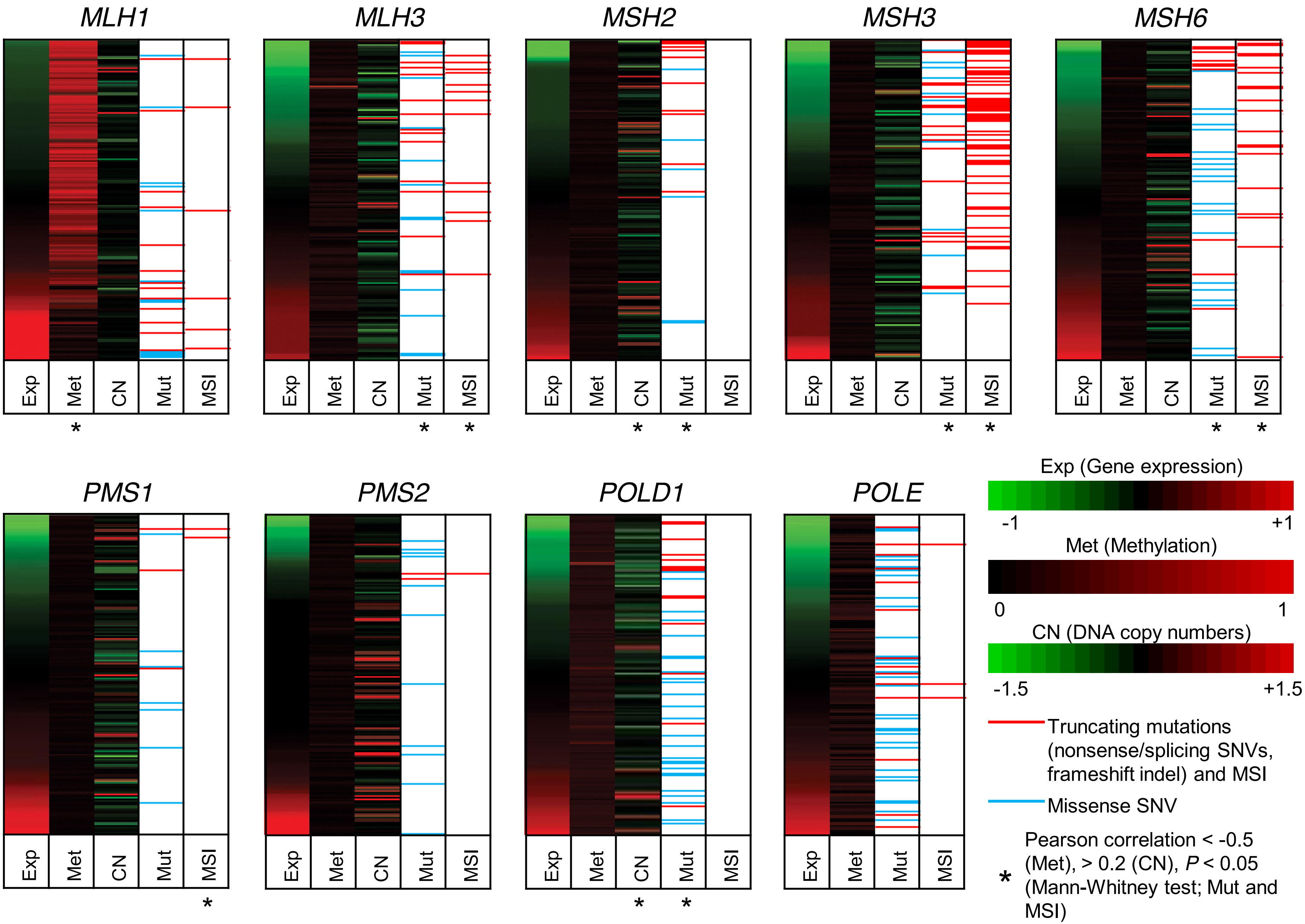

a
% of patients
0
1
2
3
4
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
b
0
10
20
30
40
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
c
% of patients
0.0
2.5
5.0
7.5
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
d
0
20
40
60
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
e
% of patients
0.0
0.5
1.0
1.5
2.0
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
f
0
10
20
30
MLH1: msi–h
MLH1: msi–l
MLH1: mss
MLH3: msi–h
MLH3: msi–l
MLH3: mss
MSH2: msi–h
MSH2: msi–l
MSH2: mss
MSH3: msi–h
MSH3: msi–l
MSH3: mss
MSH6: msi–h
MSH6: msi–l
MSH6: mss
PMS1: msi–h
PMS1: msi–l
PMS1: mss
PMS2: msi–h
PMS2: msi–l
PMS2: mss
POLD1: msi–h
POLD1: msi–l
POLD1: mss
POLE: msi–h
POLE: msi–l
POLE: mss
Frameshift
Missense
Nonsense
Splice site

Figure S7



MLH1
Exp
Met
CN
Mut
MSI
*
MLH3
Exp
Met
CN
Mut
MSI
*
*
MSH2
Exp
Met
CN
Mut
MSI
*
*
MSH3
Exp
Met
CN
Mut
MSI
*
*
MSH6
Exp
Met
CN
Mut
MSI
*
*
PMS1
Exp
Met
CN
Mut
MSI
*
PMS2
Exp
Met
CN
Mut
MSI
POLD1
Exp
Met
CN
Mut
MSI
*
*
POLE
Exp
Met
Mut
MSI
Exp (Gene expression)
-1
+1
Met (Methylation)
0
1
CN (DNA copy numbers)
-1.5
+1.5
Truncating mutations (nonsense/splicing SNVs, frameshift indel) and MSI
Missense SNV
* Pearson correlation < -0.5 (Met), > 0.2 (CN), *P* < 0.05 (Mann-Whitney test; Mut and MSI)

Figure S8

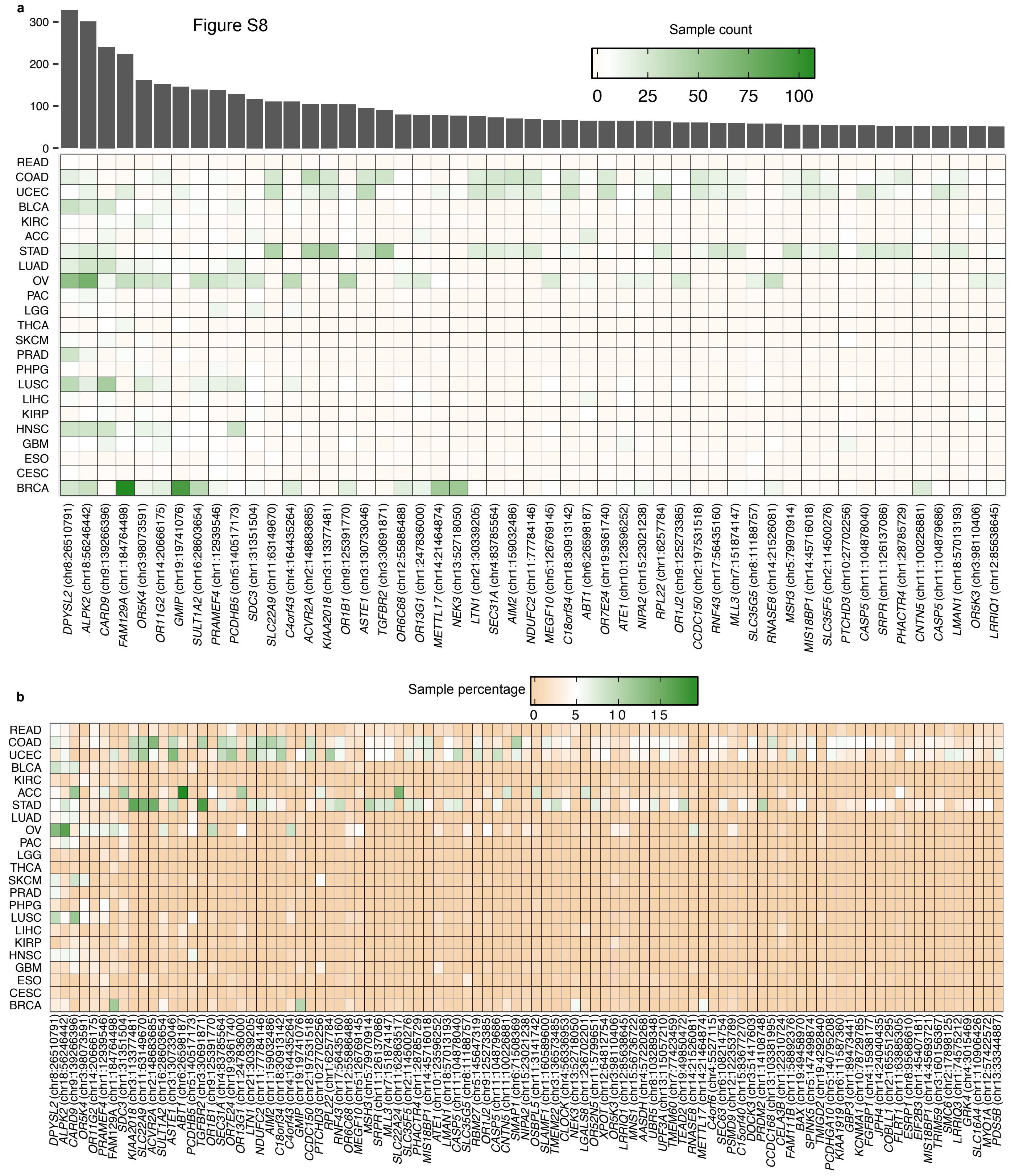

a
300
200
100
0
Sample count
0
25
50
75
100
READ
COAD
UCEC
BLCA
KIRC
ACC
STAD
LUAD
OV
PAC
LGG
THCA
SKCM
PRAD
PHPG
LUSC
LIHC
KIRP
HNSC
GBM
ESO
CESC
BRCA
DPYSL2 (chr8:26510791)
ALPK2 (chr18:56246442)
CARD9 (chr9:139266396)
FAM129A (chr1:184764498)
OR5K4 (chr3:98073591)
OR11G2 (chr14:20666175)
GMIP (chr19:19741076)
SULT1A2 (chr16:28603654)
PRAMEF4 (chr1:12939546)
PCDHB5 (chr5:140517173)
SDC3 (chr1:31351504)
SLC22A9 (chr11:63149670)
C4orf43 (chr4:164435264)
ACVR2A (chr2:148683685)
KIAA2018 (chr3:113377481)
OR1B1 (chr9:125391770)
ASTE1 (chr3:130733046)
TGFBR2 (chr3:30691871)
OR6C68 (chr12:55886488)
OR13G1 (chr1:247836000)
METTL17 (chr14:21464874)
NEK3 (chr13:52718050)
LTN1 (chr21:30339205)
SEC31A (chr4:83785564)
AIM2 (chr1:159032486)
NDUFC2 (chr11:77784146)
MEGF10 (chr5:126769145)
C18orf34 (chr18:30913142)
ABT1 (chr6:26598187)
OR7E24 (chr19:9361740)
ATE1 (chr10:123596252)
NIPA2 (chr15:23021238)
RPL22 (chr1:6257784)
OR1J2 (chr9:125273385)
CCDC150 (chr2:197531518)
RNF43 (chr17:56435160)
MLL3 (chr7:151874147)
SLC35G5 (chr8:11188757)
RNASE8 (chr14:21526081)
MSH3 (chr5:79970914)
MIS18BP1 (chr14:45716018)
SLC35F5 (chr2:114500276)
PTCHD3 (chr10:27702256)
CASP5 (chr11:104878040)
SRPR (chr11:126137086)
PHACTR4 (chr1:28785729)
CNTN5 (chr11:100226881)
CASP5 (chr11:104879686)
LMAN1 (chr18:57013193)
OR5K3 (chr3:98110406)
LRRIQ1 (chr12:85638645)
b
Sample percentage
0
5
10
15
READ
COAD
UCEC
BLCA
KIRC
ACC
STAD
LUAD
OV
PAC
LGG
THCA
SKCM
PRAD
PHPG
LUSC
LIHC
KIRP
HNSC
GBM
ESO
CESC
BRCA
DPYSL2 (chr8:26510791)
ALPK2 (chr18:56246442)
CARD9 (chr9:139266396)
OR5K4 (chr3:98073591)
OR11G2 (chr14:20666175)
PRAMEF4 (chr1:12939546)
FAM129A (chr1:184764498)
SDC3 (chr1:31351504)
KIAA2018 (chr3:113377481)
SLC22A9 (chr11:63149670)
ACVR2A (chr2:148683685)
SULT1A2 (chr16:28603654)
ASTE1 (chr3:130733046)
ABT1 (chr6:26598187)
PCDHB5 (chr5:140517173)
TGFBR2 (chr3:30691871)
OR1B1 (chr9:125391770)
SEC31A (chr4:83785564)
OR7E24 (chr19:9361740)
OR13G1 (chr1:247836000)
LTN1 (chr21:30339205)
NDUFC2 (chr11:77784146)
AIM2 (chr1:159032486)
C18orf34 (chr18:30913142)
C4orf43 (chr4:164435264)
GMIP (chr19:19741076)
CCDC150 (chr2:197531518)
PTCHD3 (chr10:27702256)
RPL22 (chr1:6257784)
RNF43 (chr17:56435160)
OR6C68 (chr12:55886488)
MEGF10 (chr5:126769145)
MSH3 (chr5:79970914)
SRPR (chr11:126137086)
MLL3 (chr7:151874147)
SLC22A24 (chr11:62863517)
SLC35F5 (chr2:114500276)
PHACTR4 (chr1:28785729)
MIS18BP1 (chr14:45716018)
ATE1 (chr10:123596252)
LMAN1 (chr18:57013193)
CASP5 (chr11:104878040)
SLC35G5 (chr8:11188757)
RBM27 (chr5:145647319)
OR1J2 (chr9:125273385)
CASP5 (chr11:104879686)
CNTN5 (chr11:100226881)
SMAP1 (chr6:71508369)
NIPA2 (chr15:23021238)
OSBPL5 (chr11:3141742)
SLAMF1 (chr1:160589600)
TMEM22 (chr3:136573485)
CLOCK (chr4:56336953)
NEK3 (chr13:52718050)
LGALS8 (chr1:236702201)
OR52N5 (chr11:5799651)
XPOT (chr12:64812754)
OR5K3 (chr3:98110406)
LRRIQ1 (chr12:85638645)
MNS1 (chr15:56736722)
AASDH (chr4:57220268)
UBR5 (chr8:103289348)
UPF3A (chr13:115057210)
TMEM60 (chr7:77423459)
TEAD2 (chr19:49850472)
RNASE8 (chr14:21526081)
METTL17 (chr14:21464874)
C4orf6 (chr4:5527115)
SEC63 (chr6:108214754)
PSMD9 (chr12:122353789)
C15orf40 (chr15:83677270)
DOCK3 (chr3:51417603)
PRDM2 (chr1:14108748)
CCDC168 (chr13:103381995)
CELA3B (chr1:22310724)
FAM111B (chr11:58892376)
BAX (chr19:49458970)
SPINK5 (chr5:147499874)
TMIGD2 (chr19:4292840)
PCDHGA10 (chr5:140795208)
KIAA1919 (chr6:111587360)
GBP3 (chr1:89473441)
KCNMA1 (chr10:78729785)
FGFBP1 (chr4:15938177)
JPH4 (chr14:24040435)
COBLL1 (chr2:165551295)
FLRT1 (chr11:63885305)
ESRP1 (chr8:95686610)
EIF2B3 (chr1:45407181)
MIS18BP1 (chr14:45693721)
TRIM59 (chr3:160156367)
SMC6 (chr2:17898125)
LRRIQ3 (chr1:74575212)
GRK4 (chr4:3015469)
SLC16A4 (chr1:110906426)
MYO1A (chr12:57422572)
PDS5B (chr13:33344887)

Figure S9

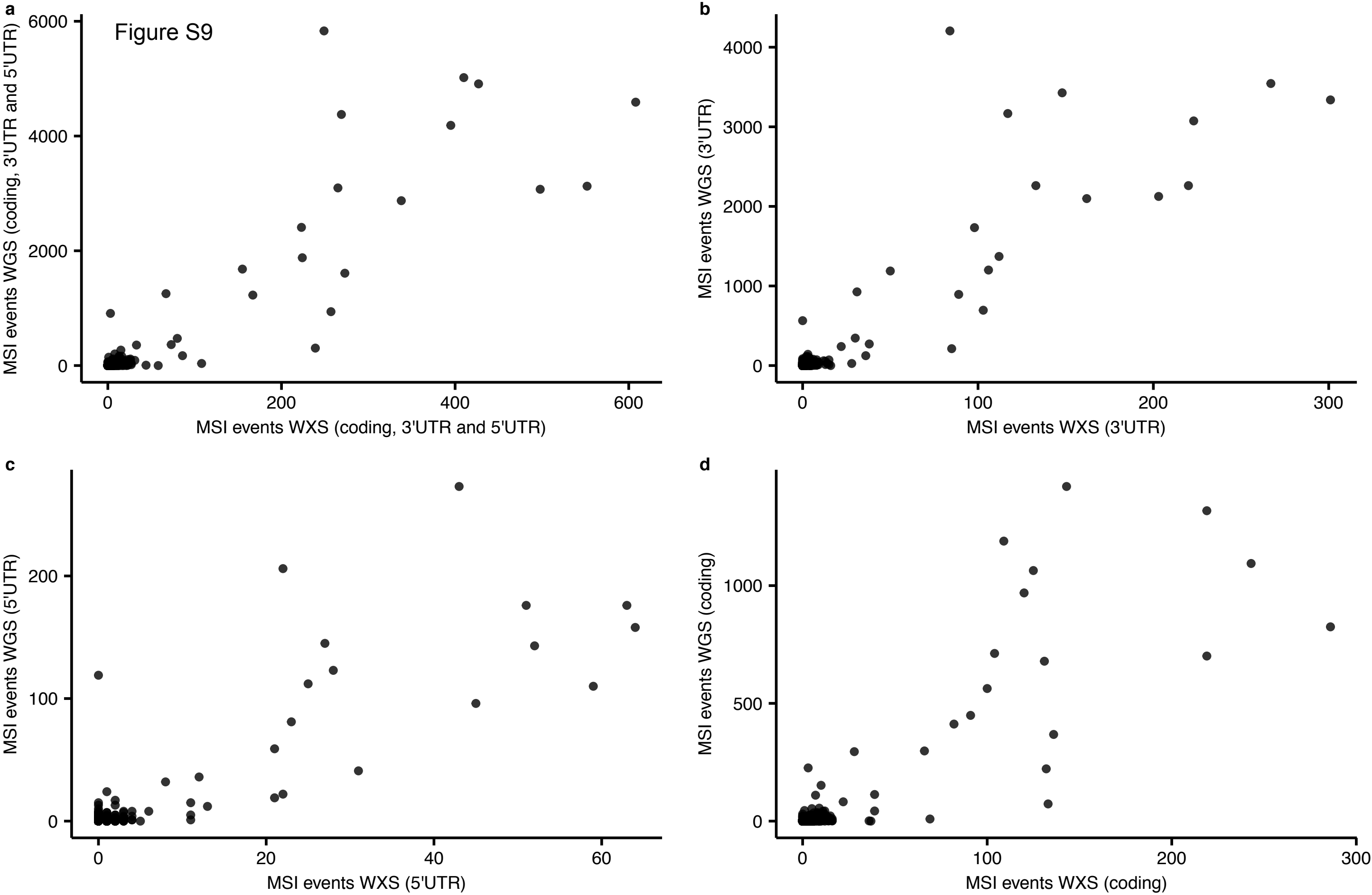
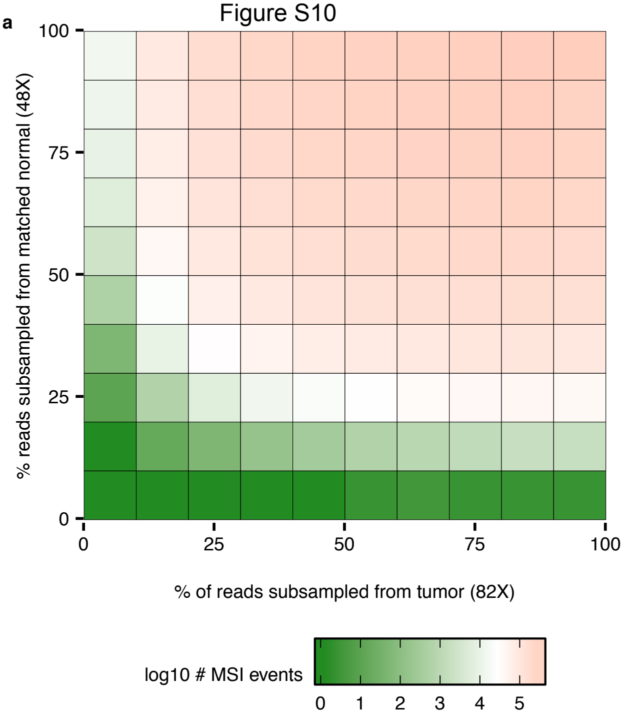
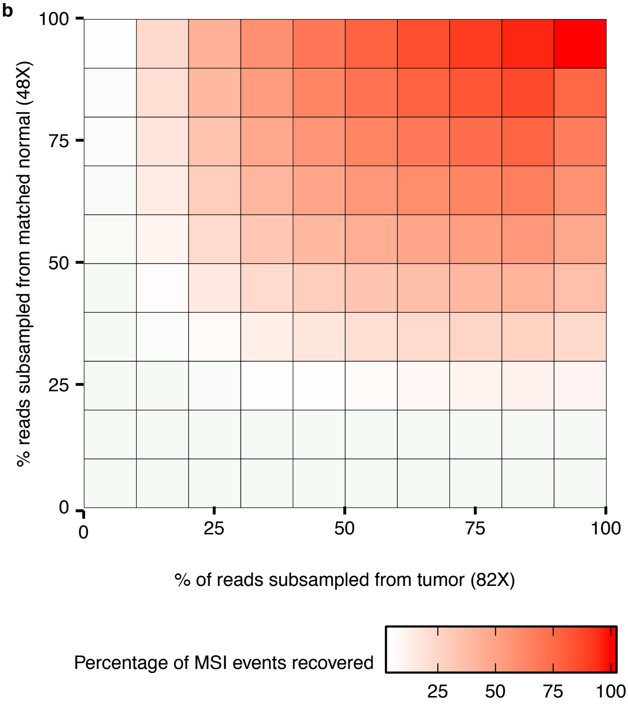

a
6000
4000
2000
0
MSI events WGS (coding, 3'UTR and 5'UTR)
0
200
400
600
MSI events WXS (coding, 3'UTR and 5'UTR)
b
4000
3000
2000
1000
0
MSI events WGS (3'UTR)
0
100
200
300
MSI events WXS (3'UTR)
c
200
100
0
MSI events WGS (5'UTR)
0
20
40
60
MSI events WXS (5'UTR)
d
1000
500
0
MSI events WGS (coding)
0
100
200
300
MSI events WXS (coding)

Figure S10


a
% reads subsampled from matched normal (48X)
100
75
50
25
0
0
25
50
75
100
% of reads subsampled from tumor (82X)
log10 # MSI events
0
1
2
3
4
5


b
% reads subsampled from matched normal (48X)
100
75
50
25
0
0
25
50
75
100
% of reads subsampled from tumor (82X)
Percentage of MSI events recovered
25
50
75
100

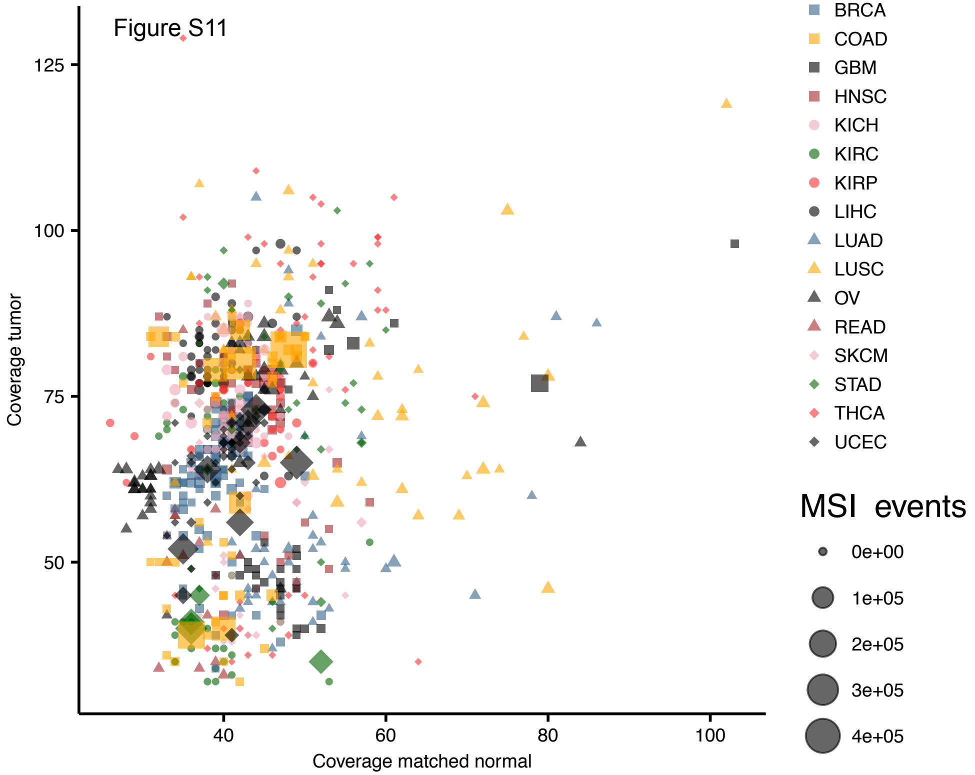
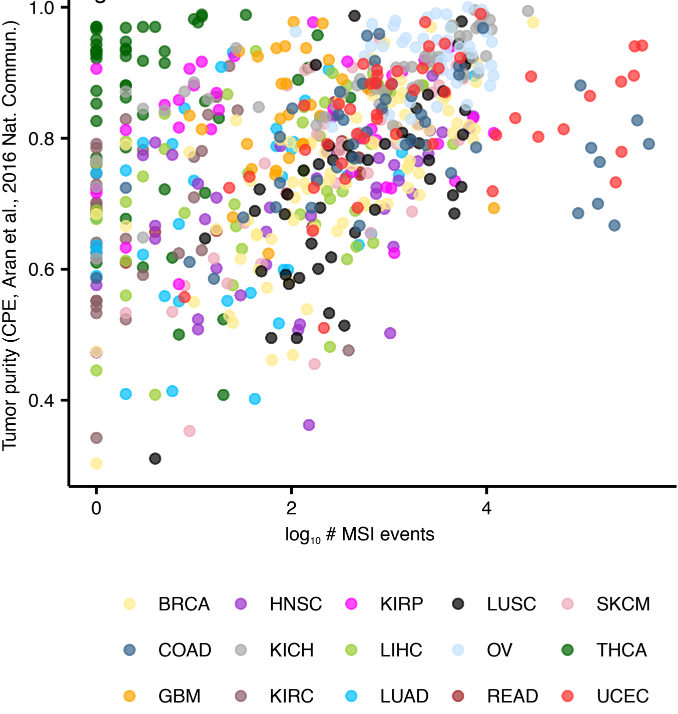


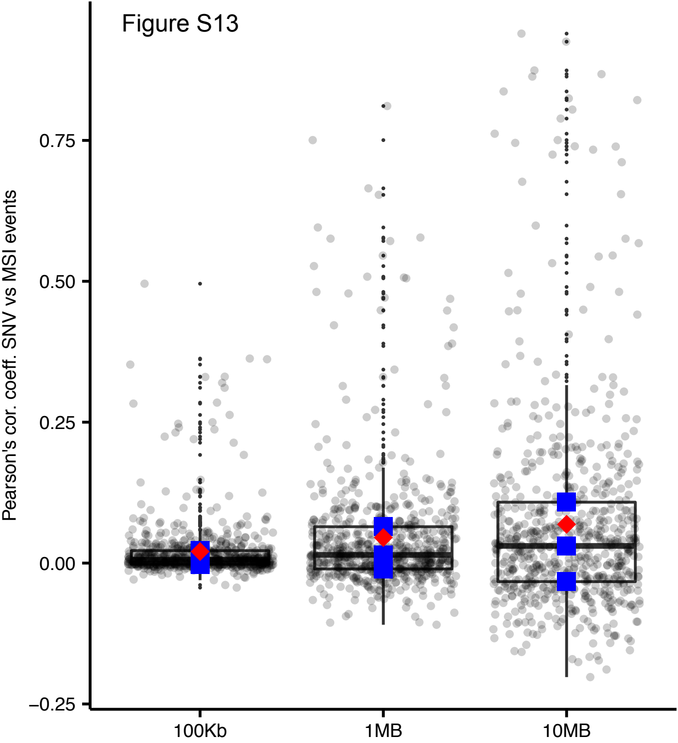
Figure S11
Coverage tumor
125
100
75
50
Coverage matched normal
40
60
80
100
BRCA
COAD
GBM
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
OV
READ
SKCM
STAD
THCA
UCEC
MSI events
0e+00
1e+05
2e+05
3e+05
4e+05

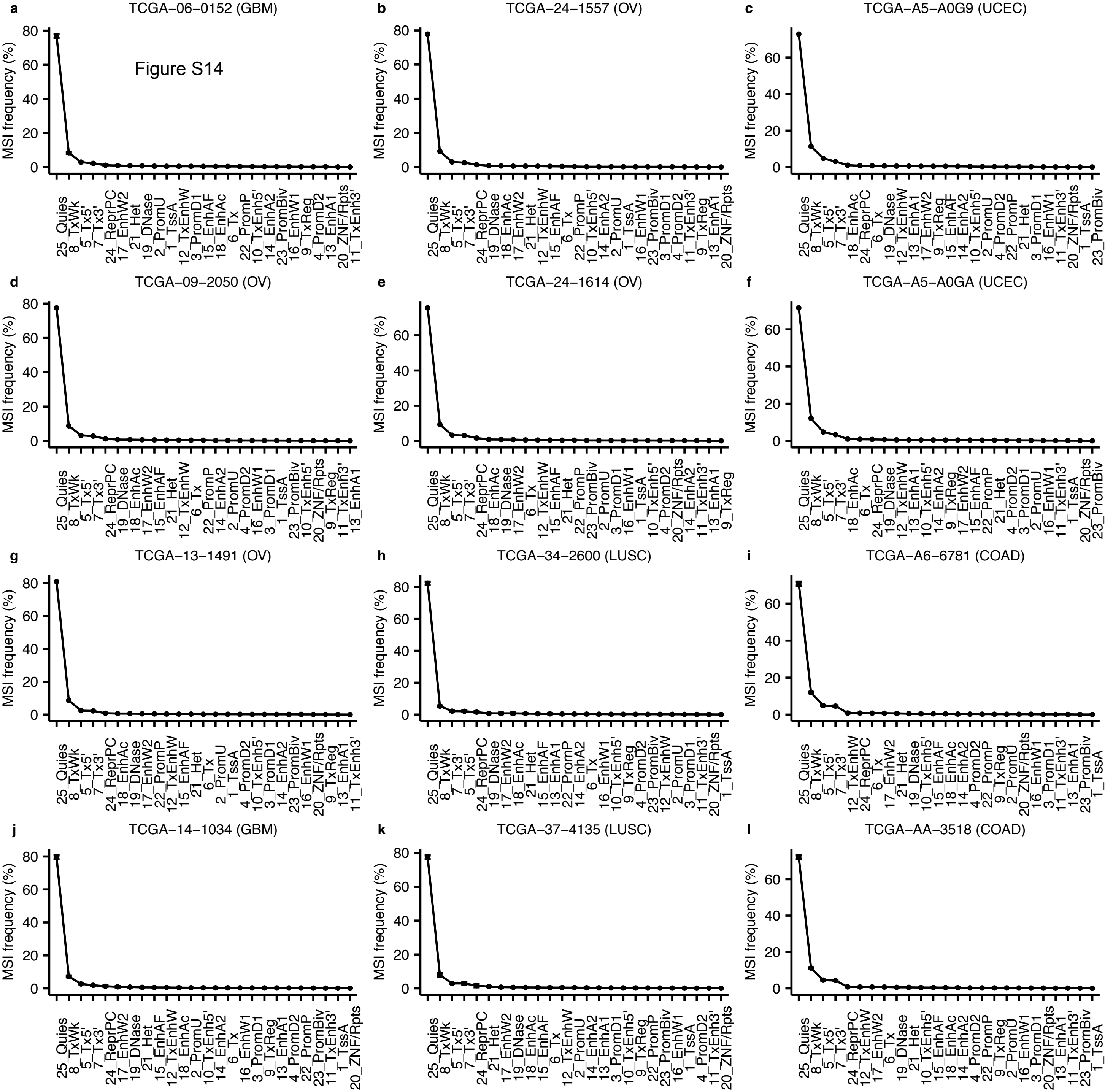
Figure S12



Tumor purity (CPE, Aran et al., 2016 Nat. Commun.)
1.0
0.8
0.6
0.4
0
2
4
$\log_{10}$ # MSI events
BRCA
HNSC
KIRP
LUSC
SKCM
COAD
KICH
LIHC
OV
THCA
GBM
KIRC
LUAD
READ
UCEC

Figure S13

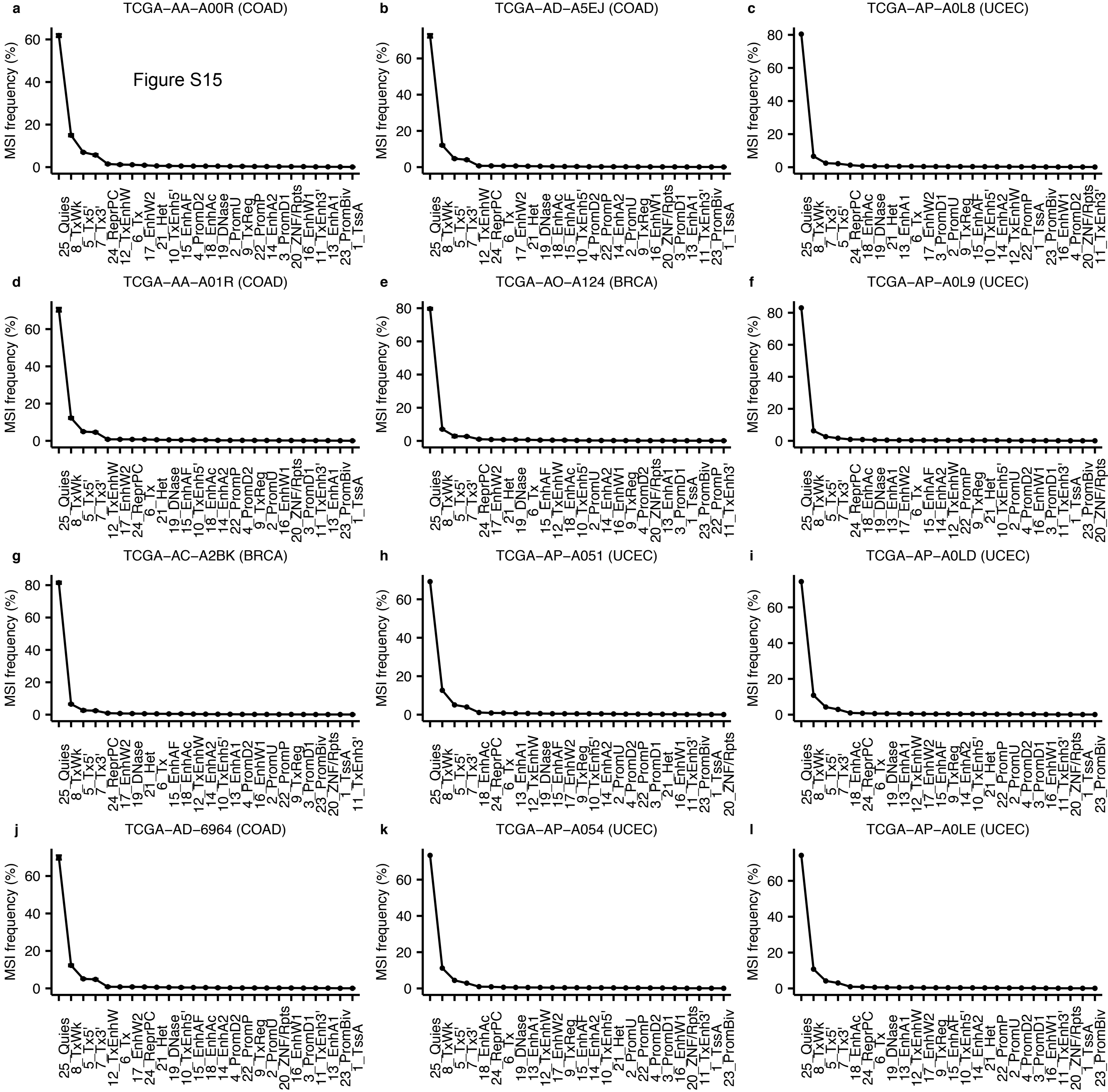

Pearson's cor. coeff. SNV vs MSI events
0.75
0.50
0.25
0.00
−0.25
100Kb
1MB
10MB



a
TCGA−06−0152 (GBM)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
17_EnhW2
21_Het
19_DNase
2_PromU
1_TssA
12_TxEnhW
3_PromD1
15_EnhAF
18_EnhAc
6_Tx
22_PromP
10_TxEnh5'
14_EnhA2
23_PromBiv
16_EnhW1
9_TxReg
4_PromD2
13_EnhA1
20_ZNF/Rpts
11_TxEnh3'
b
TCGA−24−1557 (OV)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
19_DNase
18_EnhAc
17_EnhW2
21_Het
12_TxEnhW
15_EnhAF
6_Tx
22_PromP
10_TxEnh5'
14_EnhA2
2_PromU
1_TssA
16_EnhW1
23_PromBiv
3_PromD1
4_PromD2
11_TxEnh3'
9_TxReg
13_EnhA1
20_ZNF/Rpts
c
TCGA−A5−A0G9 (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
12_TxEnhW
13_EnhA1
17_EnhW2
9_TxReg
15_EnhAF
14_EnhA2
10_TxEnh5'
2_PromU
4_PromD2
22_PromP
21_Het
3_PromD1
16_EnhW1
11_TxEnh3'
20_ZNF/Rpts
1_TssA
23_PromBiv
d
TCGA−09−2050 (OV)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
19_DNase
18_EnhAc
17_EnhW2
15_EnhAF
21_Het
12_TxEnhW
6_Tx
22_PromP
14_EnhA2
2_PromU
4_PromD2
16_EnhW1
3_PromD1
1_TssA
23_PromBiv
10_TxEnh5'
20_ZNF/Rpts
9_TxReg
11_TxEnh3'
13_EnhA1
e
TCGA−24−1614 (OV)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
18_EnhAc
19_DNase
17_EnhW2
6_Tx
12_TxEnhW
15_EnhAF
21_Het
22_PromP
23_PromBiv
2_PromU
3_PromD1
16_EnhW1
1_TssA
10_TxEnh5'
4_PromD2
20_ZNF/Rpts
14_EnhA2
11_TxEnh3'
13_EnhA1
9_TxReg
f
TCGA−A5−A0GA (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
6_Tx
24_ReprPC
19_DNase
12_TxEnhW
13_EnhA1
10_TxEnh5'
14_EnhA2
9_TxReg
17_EnhW2
15_EnhAF
22_PromP
21_Het
4_PromD2
3_PromD1
2_PromU
16_EnhW1
11_TxEnh3'
1_TssA
20_ZNF/Rpts
23_PromBiv
g
TCGA−13−1491 (OV)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
18_EnhAc
19_DNase
17_EnhW2
22_PromP
12_TxEnhW
15_EnhAF
21_Het
6_Tx
2_PromU
1_TssA
4_PromD2
10_TxEnh5'
3_PromD1
14_EnhA2
23_PromBiv
16_EnhW1
20_ZNF/Rpts
9_TxReg
13_EnhA1
11_TxEnh3'
h
TCGA−34−2600 (LUSC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
19_DNase
17_EnhW2
18_EnhAc
21_Het
15_EnhAF
13_EnhA1
22_PromP
14_EnhA2
6_Tx
16_EnhW1
10_TxEnh5'
9_TxReg
4_PromD2
23_PromBiv
12_TxEnhW
2_PromU
3_PromD1
11_TxEnh3'
20_ZNF/Rpts
1_TssA
i
TCGA−A6−6781 (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
12_TxEnhW
24_ReprPC
6_Tx
17_EnhW2
21_Het
19_DNase
10_TxEnh5'
15_EnhAF
18_EnhAc
14_EnhA2
4_PromD2
22_PromP
9_TxReg
2_PromU
20_ZNF/Rpts
16_EnhW1
3_PromD1
11_TxEnh3'
13_EnhA1
23_PromBiv
1_TssA
j
TCGA−14−1034 (GBM)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
17_EnhW2
19_DNase
21_Het
15_EnhAF
12_TxEnhW
18_EnhAc
2_PromU
10_TxEnh5'
14_EnhA2
6_Tx
16_EnhW1
3_PromD1
9_TxReg
13_EnhA1
4_PromD2
22_PromP
23_PromBiv
11_TxEnh3'
1_TssA
20_ZNF/Rpts
k
TCGA−37−4135 (LUSC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
21_Het
17_EnhW2
19_DNase
18_EnhAc
15_EnhAF
6_Tx
12_TxEnhW
2_PromU
14_EnhA2
13_EnhA1
3_PromD1
10_TxEnh5'
9_TxReg
22_PromP
23_PromBiv
16_EnhW1
1_TssA
4_PromD2
11_TxEnh3'
20_ZNF/Rpts
l
TCGA−AA−3518 (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
12_TxEnhW
17_EnhW2
6_Tx
19_DNase
21_Het
10_TxEnh5'
15_EnhAF
18_EnhAc
14_EnhA2
4_PromD2
22_PromP
9_TxReg
2_PromU
16_EnhW1
3_PromD1
20_ZNF/Rpts
13_EnhA1
11_TxEnh3'
23_PromBiv
1_TssA

Figure S14

Figure S15


a
TCGA–AA–A00R (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
12_TxEnhW
6_Tx
17_EnhW2
21_Het
10_TxEnh5'
15_EnhAF
4_PromD2
18_EnhAc
19_DNase
2_PromU
9_TxReg
22_PromP
14_EnhA2
3_PromD1
20_ZNF/Rpts
16_EnhW1
11_TxEnh3'
13_EnhA1
23_PromBiv
1_TssA
b
TCGA–AD–A5EJ (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
12_TxEnhW
24_ReprPC
6_Tx
17_EnhW2
21_Het
19_DNase
18_EnhAc
15_EnhAF
10_TxEnh5'
4_PromD2
22_PromP
14_EnhA2
2_PromU
9_TxReg
16_EnhW1
20_ZNF/Rpts
3_PromD1
13_EnhA1
11_TxEnh3'
23_PromBiv
1_TssA
c
TCGA–AP–A0L8 (UCEC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
18_EnhAc
19_DNase
21_Het
13_EnhA1
6_Tx
17_EnhW2
3_PromD1
2_PromU
9_TxReg
15_EnhAF
10_TxEnh5'
14_EnhA2
12_TxEnhW
22_PromP
1_TssA
23_PromBiv
16_EnhW1
4_PromD2
20_ZNF/Rpts
11_TxEnh3'
d
TCGA–AA–A01R (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
12_TxEnhW
17_EnhW2
24_ReprPC
6_Tx
21_Het
19_DNase
15_EnhAF
10_TxEnh5'
18_EnhAc
14_EnhA2
22_PromP
4_PromD2
9_TxReg
2_PromU
16_EnhW1
20_ZNF/Rpts
3_PromD1
11_TxEnh3'
13_EnhA1
23_PromBiv
1_TssA
e
TCGA–AO–A124 (BRCA)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
17_EnhW2
21_Het
19_DNase
6_Tx
15_EnhAF
12_TxEnhW
18_EnhAc
10_TxEnh5'
2_PromU
14_EnhA2
16_EnhW1
9_TxReg
4_PromD2
20_ZNF/Rpts
13_EnhA1
3_PromD1
1_TssA
23_PromBiv
22_PromP
11_TxEnh3'
f
TCGA–AP–A0L9 (UCEC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
18_EnhAc
19_DNase
13_EnhA1
17_EnhW2
6_Tx
15_EnhAF
14_EnhA2
12_TxEnhW
22_PromP
9_TxReg
21_Het
10_TxEnh5'
2_PromU
4_PromD2
16_EnhW1
3_PromD1
11_TxEnh3'
1_TssA
23_PromBiv
20_ZNF/Rpts
g
TCGA–AC–A2BK (BRCA)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
17_EnhW2
19_DNase
21_Het
6_Tx
15_EnhAF
18_EnhAc
12_TxEnhW
14_EnhA2
10_TxEnh5'
13_EnhA1
4_PromD2
16_EnhW1
2_PromU
22_PromP
9_TxReg
3_PromD1
23_PromBiv
20_ZNF/Rpts
1_TssA
11_TxEnh3'
h
TCGA–AP–A051 (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
13_EnhA1
12_TxEnhW
19_DNase
15_EnhAF
17_EnhW2
9_TxReg
10_TxEnh5'
14_EnhA2
2_PromU
4_PromD2
22_PromP
3_PromD1
21_Het
16_EnhW1
11_TxEnh3'
23_PromBiv
1_TssA
20_ZNF/Rpts
i
TCGA–AP–A0LD (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
13_EnhA1
12_TxEnhW
17_EnhW2
15_EnhAF
9_TxReg
14_EnhA2
10_TxEnh5'
21_Het
22_PromP
2_PromU
4_PromD2
3_PromD1
16_EnhW1
11_TxEnh3'
20_ZNF/Rpts
1_TssA
23_PromBiv
j
TCGA–AD–6964 (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
12_TxEnhW
6_Tx
17_EnhW2
24_ReprPC
21_Het
19_DNase
10_TxEnh5'
15_EnhAF
18_EnhAc
14_EnhA2
4_PromD2
22_PromP
9_TxReg
2_PromU
20_ZNF/Rpts
16_EnhW1
3_PromD1
11_TxEnh3'
13_EnhA1
23_PromBiv
1_TssA
k
TCGA–AP–A054 (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
13_EnhA1
12_TxEnhW
17_EnhW2
9_TxReg
15_EnhAF
14_EnhA2
10_TxEnh5'
21_Het
2_PromU
22_PromP
4_PromD2
3_PromD1
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
23_PromBiv
l
TCGA–AP–A0LE (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
13_EnhA1
12_TxEnhW
17_EnhW2
9_TxReg
15_EnhAF
10_TxEnh5'
14_EnhA2
21_Het
22_PromP
2_PromU
4_PromD2
3_PromD1
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
23_PromBiv

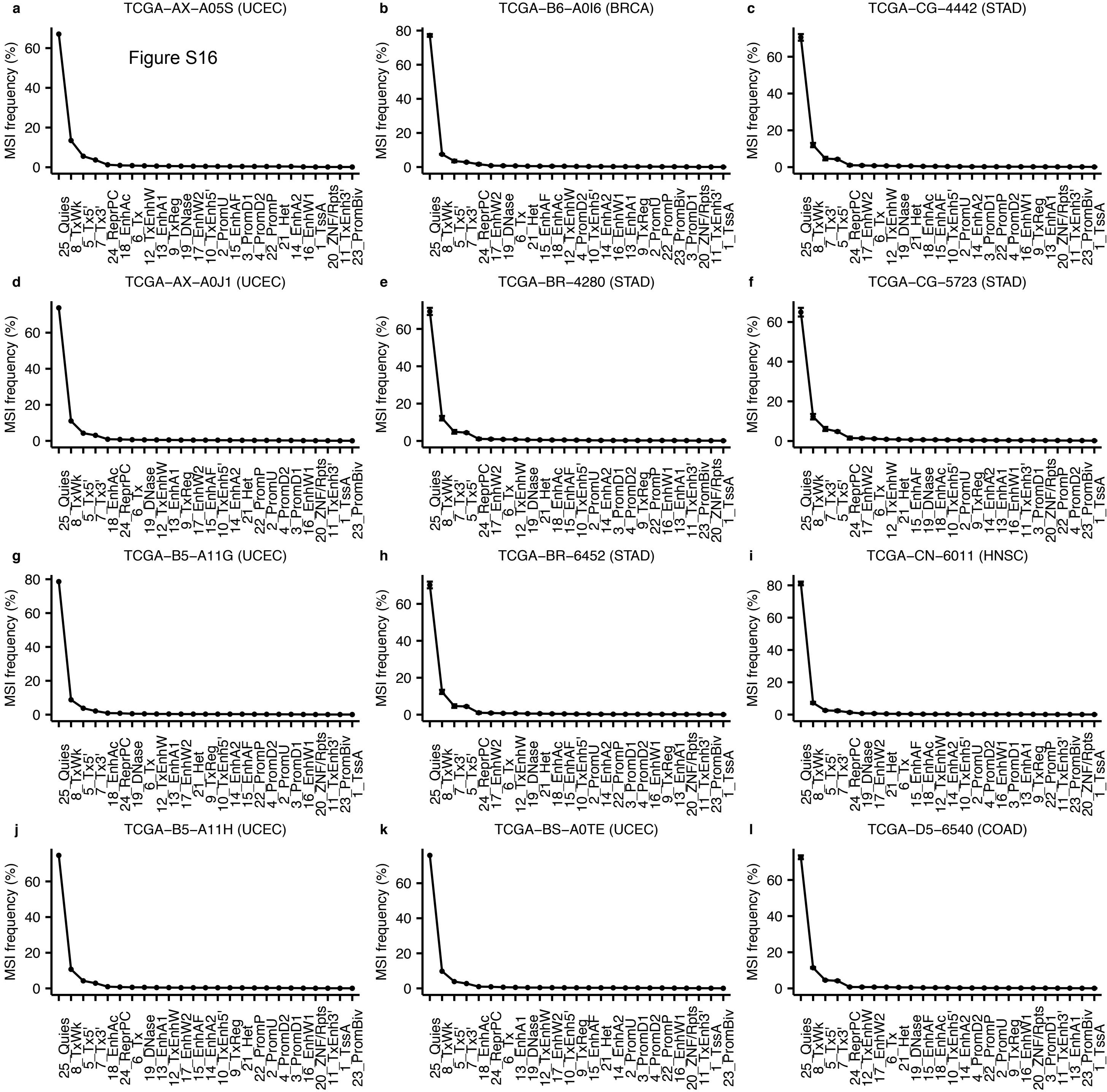


Figure S16
a
TCGA−AX−A05S (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
18_EnhAc
6_Tx
12_TxEnhW
13_EnhA1
9_TxReg
19_DNase
17_EnhW2
10_TxEnh5'
2_PromU
15_EnhAF
3_PromD1
4_PromD2
22_PromP
21_Het
14_EnhA2
16_EnhW1
1_TssA
20_ZNF/Rpts
11_TxEnh3'
23_PromBiv
b
TCGA−B6−A0I6 (BRCA)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
17_EnhW2
19_DNase
6_Tx
21_Het
15_EnhAF
18_EnhAc
12_TxEnhW
4_PromD2
10_TxEnh5'
14_EnhA2
16_EnhW1
13_EnhA1
9_TxReg
2_PromU
22_PromP
23_PromBiv
3_PromD1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
c
TCGA−CG−4442 (STAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
6_Tx
12_TxEnhW
19_DNase
21_Het
18_EnhAc
15_EnhAF
10_TxEnh5'
2_PromU
14_EnhA2
3_PromD1
22_PromP
4_PromD2
16_EnhW1
9_TxReg
13_EnhA1
20_ZNF/Rpts
11_TxEnh3'
23_PromBiv
1_TssA
d
TCGA−AX−A0J1 (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
12_TxEnhW
13_EnhA1
9_TxReg
17_EnhW2
15_EnhAF
10_TxEnh5'
14_EnhA2
21_Het
22_PromP
2_PromU
4_PromD2
3_PromD1
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
23_PromBiv
e
TCGA−BR−4280 (STAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
6_Tx
12_TxEnhW
19_DNase
21_Het
18_EnhAc
15_EnhAF
10_TxEnh5'
2_PromU
14_EnhA2
3_PromD1
4_PromD2
9_TxReg
22_PromP
16_EnhW1
13_EnhA1
11_TxEnh3'
23_PromBiv
20_ZNF/Rpts
1_TssA
f
TCGA−CG−5723 (STAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
6_Tx
12_TxEnhW
21_Het
15_EnhAF
19_DNase
18_EnhAc
10_TxEnh5'
2_PromU
9_TxReg
14_EnhA2
13_EnhA1
16_EnhW1
11_TxEnh3'
3_PromD1
20_ZNF/Rpts
22_PromP
4_PromD2
23_PromBiv
1_TssA
g
TCGA−B5−A11G (UCEC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
19_DNase
6_Tx
12_TxEnhW
13_EnhA1
17_EnhW2
21_Het
9_TxReg
10_TxEnh5'
14_EnhA2
15_EnhAF
22_PromP
4_PromD2
2_PromU
3_PromD1
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
23_PromBiv
1_TssA
h
TCGA−BR−6452 (STAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
6_Tx
12_TxEnhW
19_DNase
21_Het
18_EnhAc
15_EnhAF
10_TxEnh5'
2_PromU
14_EnhA2
22_PromP
3_PromD1
4_PromD2
16_EnhW1
9_TxReg
13_EnhA1
20_ZNF/Rpts
11_TxEnh3'
23_PromBiv
1_TssA
i
TCGA−CN−6011 (HNSC)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
19_DNase
17_EnhW2
21_Het
6_Tx
15_EnhAF
18_EnhAc
12_TxEnhW
14_EnhA2
10_TxEnh5'
2_PromU
4_PromD2
16_EnhW1
3_PromD1
13_EnhA1
9_TxReg
22_PromP
11_TxEnh3'
23_PromBiv
20_ZNF/Rpts
1_TssA
j
TCGA−B5−A11H (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
19_DNase
13_EnhA1
12_TxEnhW
17_EnhW2
15_EnhAF
14_EnhA2
10_TxEnh5'
9_TxReg
21_Het
22_PromP
2_PromU
4_PromD2
3_PromD1
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
23_PromBiv
k
TCGA−BS−A0TE (UCEC)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
6_Tx
13_EnhA1
19_DNase
12_TxEnhW
17_EnhW2
10_TxEnh5'
9_TxReg
15_EnhAF
21_Het
14_EnhA2
2_PromU
3_PromD1
4_PromD2
22_PromP
16_EnhW1
20_ZNF/Rpts
11_TxEnh3'
1_TssA
23_PromBiv
l
TCGA−D5−6540 (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
24_ReprPC
12_TxEnhW
17_EnhW2
6_Tx
21_Het
19_DNase
15_EnhAF
18_EnhAc
10_TxEnh5'
14_EnhA2
4_PromD2
22_PromP
2_PromU
9_TxReg
16_EnhW1
20_ZNF/Rpts
3_PromD1
11_TxEnh3'
13_EnhA1
23_PromBiv
1_TssA

Figure S17

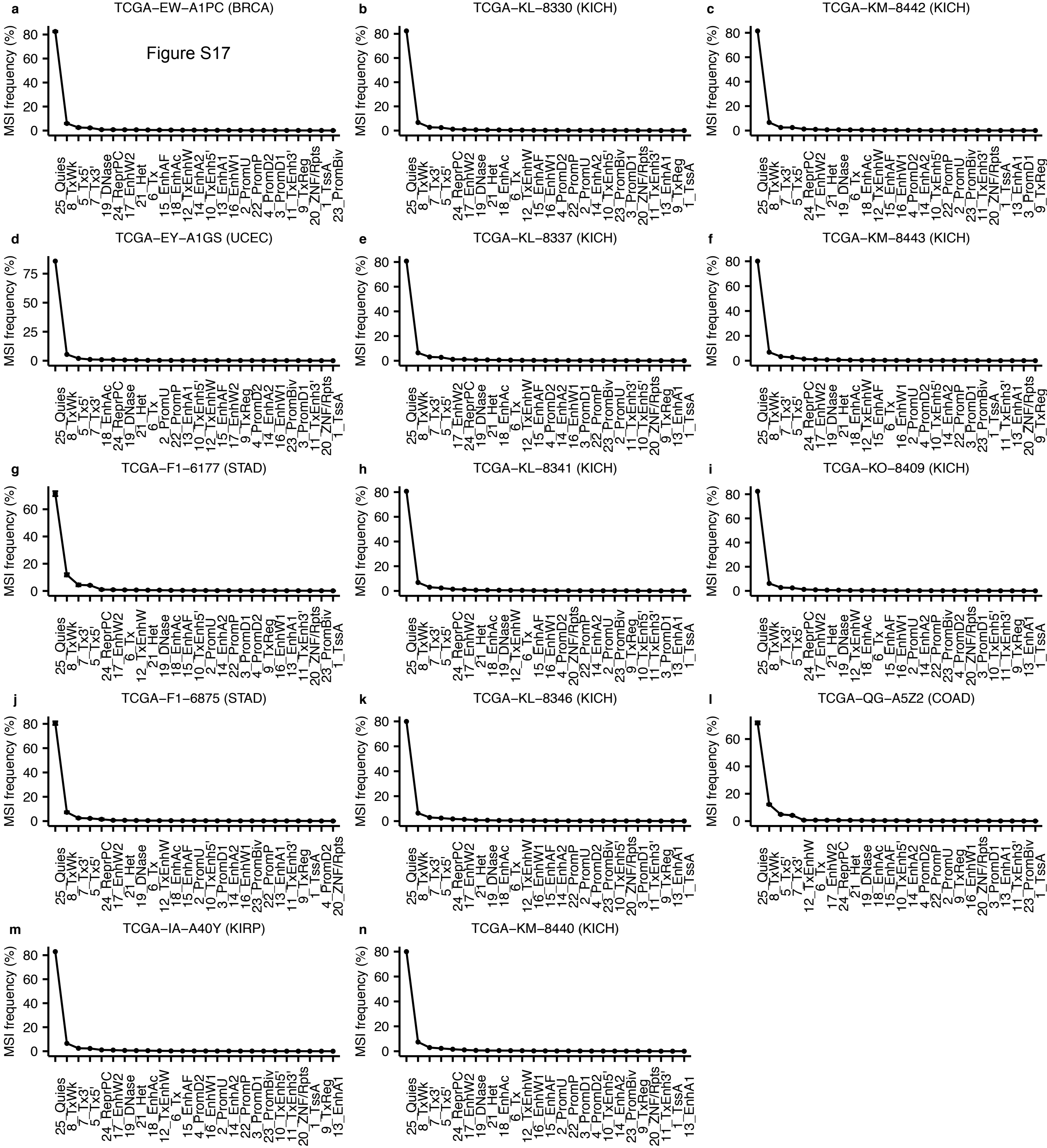

a
TCGA−EW−A1PC (BRCA)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
19_DNase
24_ReprPC
17_EnhW2
21_Het
6_Tx
15_EnhAF
18_EnhAc
12_TxEnhW
14_EnhA2
10_TxEnh5'
13_EnhA1
16_EnhW1
2_PromU
22_PromP
4_PromD2
3_PromD1
11_TxEnh3'
9_TxReg
20_ZNF/Rpts
1_TssA
23_PromBiv
b
TCGA−KL−8330 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
19_DNase
21_Het
18_EnhAc
6_Tx
12_TxEnhW
15_EnhAF
16_EnhW1
4_PromD2
22_PromP
2_PromU
14_EnhA2
10_TxEnh5'
23_PromBiv
3_PromD1
20_ZNF/Rpts
11_TxEnh3'
13_EnhA1
9_TxReg
1_TssA
c
TCGA−KM−8442 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
21_Het
19_DNase
6_Tx
18_EnhAc
12_TxEnhW
15_EnhAF
16_EnhW1
4_PromD2
14_EnhA2
10_TxEnh5'
22_PromP
2_PromU
23_PromBiv
11_TxEnh3'
20_ZNF/Rpts
1_TssA
13_EnhA1
3_PromD1
9_TxReg
d
TCGA−EY−A1GS (UCEC)
MSI frequency (%)
0
25
50
75
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
18_EnhAc
24_ReprPC
19_DNase
21_Het
6_Tx
2_PromU
22_PromP
13_EnhA1
10_TxEnh5'
12_TxEnhW
15_EnhAF
17_EnhW2
9_TxReg
4_PromD2
14_EnhA2
16_EnhW1
23_PromBiv
3_PromD1
11_TxEnh3'
20_ZNF/Rpts
1_TssA
e
TCGA−KL−8337 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
17_EnhW2
24_ReprPC
19_DNase
21_Het
18_EnhAc
6_Tx
12_TxEnhW
15_EnhAF
4_PromD2
14_EnhA2
16_EnhW1
3_PromD1
22_PromP
23_PromBiv
2_PromU
11_TxEnh3'
10_TxEnh5'
20_ZNF/Rpts
9_TxReg
13_EnhA1
1_TssA
f
TCGA−KM−8443 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
19_DNase
21_Het
18_EnhAc
12_TxEnhW
15_EnhAF
6_Tx
16_EnhW1
2_PromU
4_PromD2
10_TxEnh5'
14_EnhA2
22_PromP
3_PromD1
23_PromBiv
1_TssA
11_TxEnh3'
13_EnhA1
20_ZNF/Rpts
9_TxReg
g
TCGA−F1−6177 (STAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
6_Tx
12_TxEnhW
21_Het
19_DNase
18_EnhAc
15_EnhAF
10_TxEnh5'
2_PromU
14_EnhA2
22_PromP
3_PromD1
4_PromD2
9_TxReg
16_EnhW1
13_EnhA1
11_TxEnh3'
20_ZNF/Rpts
23_PromBiv
1_TssA
h
TCGA−KL−8341 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
21_Het
18_EnhAc
19_DNase
12_TxEnhW
6_Tx
15_EnhAF
16_EnhW1
4_PromD2
20_ZNF/Rpts
22_PromP
14_EnhA2
2_PromU
23_PromBiv
9_TxReg
10_TxEnh5'
11_TxEnh3'
3_PromD1
13_EnhA1
1_TssA
i
TCGA−KO−8409 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
21_Het
19_DNase
12_TxEnhW
18_EnhAc
6_Tx
15_EnhAF
16_EnhW1
2_PromU
14_EnhA2
22_PromP
23_PromBiv
4_PromD2
20_ZNF/Rpts
3_PromD1
10_TxEnh5'
11_TxEnh3'
9_TxReg
13_EnhA1
1_TssA
j
TCGA−F1−6875 (STAD)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
21_Het
19_DNase
6_Tx
12_TxEnhW
18_EnhAc
15_EnhAF
2_PromU
10_TxEnh5'
3_PromD1
14_EnhA2
16_EnhW1
23_PromBiv
22_PromP
13_EnhA1
11_TxEnh3'
9_TxReg
1_TssA
4_PromD2
20_ZNF/Rpts
k
TCGA−KL−8346 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
21_Het
19_DNase
18_EnhAc
6_Tx
12_TxEnhW
16_EnhW1
15_EnhAF
14_EnhA2
22_PromP
2_PromU
4_PromD2
23_PromBiv
10_TxEnh5'
20_ZNF/Rpts
3_PromD1
11_TxEnh3'
9_TxReg
13_EnhA1
1_TssA
l
TCGA−QG−A5Z2 (COAD)
MSI frequency (%)
0
20
40
60
25_Quies
8_TxWk
5_Tx5'
7_Tx3'
12_TxEnhW
6_Tx
17_EnhW2
24_ReprPC
21_Het
19_DNase
18_EnhAc
15_EnhAF
10_TxEnh5'
14_EnhA2
4_PromD2
22_PromP
2_PromU
9_TxReg
16_EnhW1
20_ZNF/Rpts
3_PromD1
13_EnhA1
11_TxEnh3'
23_PromBiv
1_TssA
m
TCGA−IA−A40Y (KIRP)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
19_DNase
21_Het
18_EnhAc
12_TxEnhW
6_Tx
15_EnhAF
4_PromD2
16_EnhW1
2_PromU
14_EnhA2
22_PromP
3_PromD1
23_PromBiv
10_TxEnh5'
11_TxEnh3'
20_ZNF/Rpts
1_TssA
9_TxReg
13_EnhA1
n
TCGA−KM−8440 (KICH)
MSI frequency (%)
0
20
40
60
80
25_Quies
8_TxWk
7_Tx3'
5_Tx5'
24_ReprPC
17_EnhW2
19_DNase
21_Het
18_EnhAc
6_Tx
12_TxEnhW
16_EnhW1
15_EnhAF
2_PromU
22_PromP
3_PromD1
4_PromD2
10_TxEnh5'
14_EnhA2
23_PromBiv
9_TxReg
20_ZNF/Rpts
11_TxEnh3'
1_TssA
13_EnhA1

Figure S18

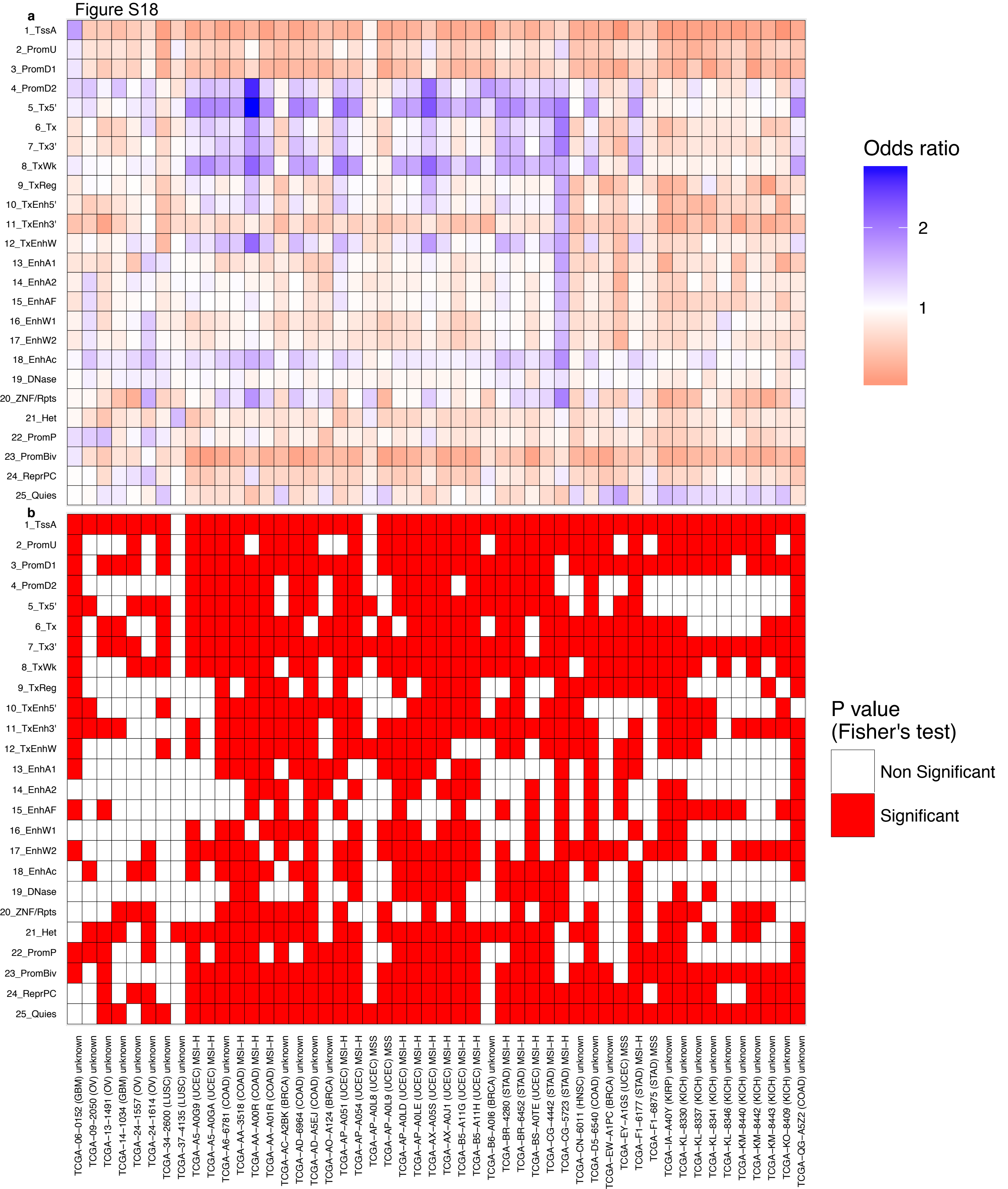
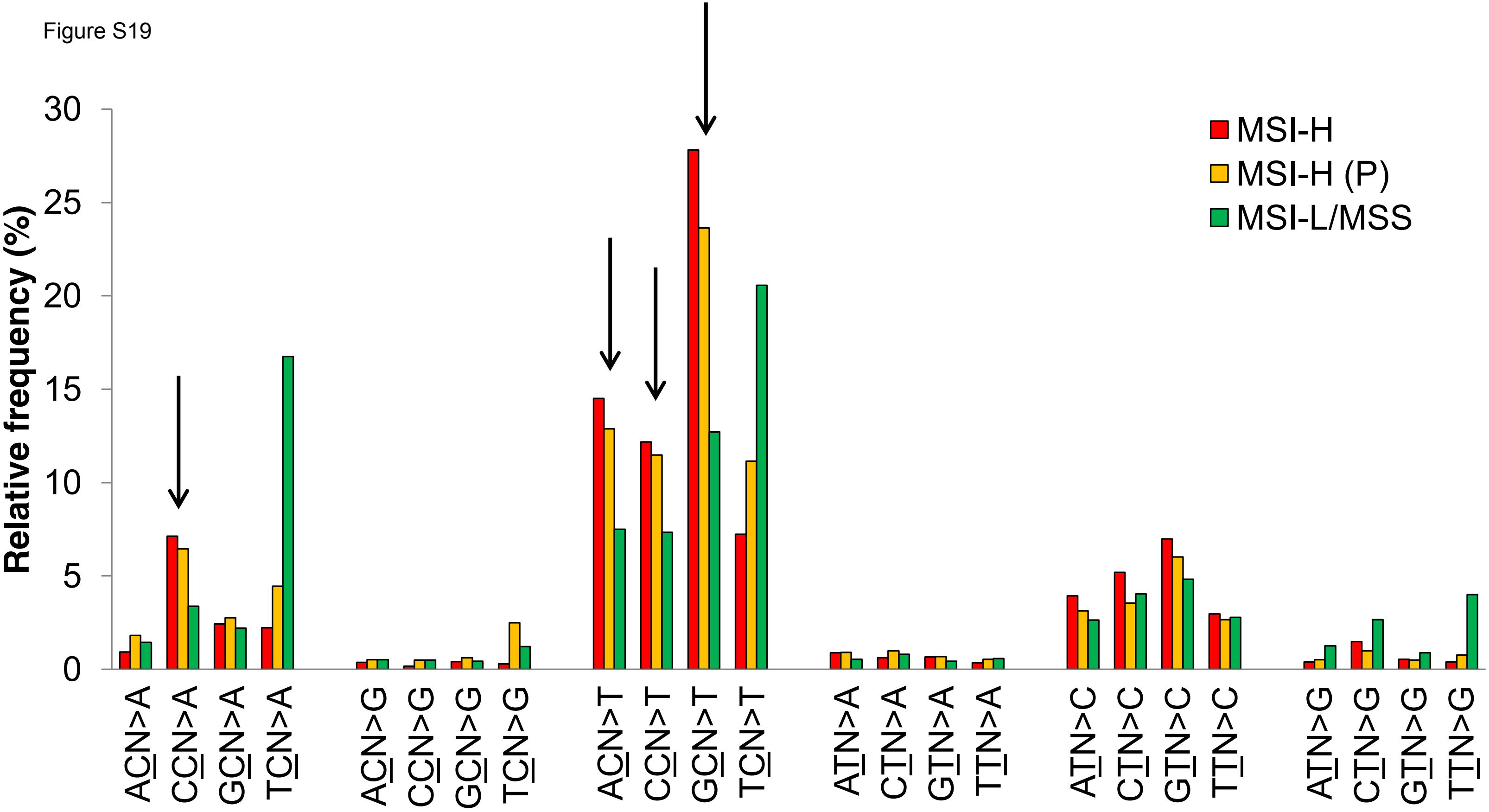

a
1_TssA
2_PromU
3_PromD1
4_PromD2
5_Tx5'
6_Tx
7_Tx3'
8_TxWk
9_TxReg
10_TxEnh5'
11_TxEnh3'
12_TxEnhW
13_EnhA1
14_EnhA2
15_EnhAF
16_EnhW1
17_EnhW2
18_EnhAc
19_DNase
20_ZNF/Rpts
21_Het
22_PromP
23_PromBiv
24_ReprPC
25_Quies
Odds ratio
2
1
b
1_TssA
2_PromU
3_PromD1
4_PromD2
5_Tx5'
6_Tx
7_Tx3'
8_TxWk
9_TxReg
10_TxEnh5'
11_TxEnh3'
12_TxEnhW
13_EnhA1
14_EnhA2
15_EnhAF
16_EnhW1
17_EnhW2
18_EnhAc
19_DNase
20_ZNF/Rpts
21_Het
22_PromP
23_PromBiv
24_ReprPC
25_Quies
P value
(Fisher's test)
Non Significant
Significant
TCGA–06–0152 (GBM) unknown
TCGA–09–2050 (OV) unknown
TCGA–13–1491 (OV) unknown
TCGA–14–1034 (GBM) unknown
TCGA–24–1557 (OV) unknown
TCGA–24–1614 (OV) unknown
TCGA–34–2600 (LUSC) unknown
TCGA–37–4135 (LUSC) unknown
TCGA–A5–A0G9 (UCEC) MSI–H
TCGA–A5–A0GA (UCEC) MSI–H
TCGA–A6–6781 (COAD) unknown
TCGA–AA–3518 (COAD) MSI–H
TCGA–AA–A00R (COAD) MSI–H
TCGA–AA–A01R (COAD) MSI–H
TCGA–AC–A2BK (BRCA) unknown
TCGA–AD–6964 (COAD) unknown
TCGA–AD–A5EJ (COAD) unknown
TCGA–AO–A124 (BRCA) unknown
TCGA–AP–A051 (UCEC) MSI–H
TCGA–AP–A054 (UCEC) MSI–H
TCGA–AP–A0L8 (UCEC) MSS
TCGA–AP–A0L9 (UCEC) MSS
TCGA–AP–A0LD (UCEC) MSI–H
TCGA–AP–A0LE (UCEC) MSI–H
TCGA–AX–A05S (UCEC) MSI–H
TCGA–AX–A0J1 (UCEC) MSI–H
TCGA–B5–A11G (UCEC) MSI–H
TCGA–B5–A11H (UCEC) MSI–H
TCGA–B6–A0I6 (BRCA) unknown
TCGA–BR–4280 (STAD) MSI–H
TCGA–BR–6452 (STAD) MSI–H
TCGA–BS–A0TE (UCEC) MSI–H
TCGA–CG–4442 (STAD) MSI–H
TCGA–CG–5723 (STAD) MSI–H
TCGA–CN–6011 (HNSC) unknown
TCGA–D5–6540 (COAD) unknown
TCGA–EW–A1PC (BRCA) unknown
TCGA–EY–A1GS (UCEC) MSS
TCGA–F1–6177 (STAD) MSI–H
TCGA–F1–6875 (STAD) MSS
TCGA–IA–A40Y (KIRP) unknown
TCGA–KL–8330 (KICH) unknown
TCGA–KL–8337 (KICH) unknown
TCGA–KL–8341 (KICH) unknown
TCGA–KL–8346 (KICH) unknown
TCGA–KM–8440 (KICH) unknown
TCGA–KM–8442 (KICH) unknown
TCGA–KM–8443 (KICH) unknown
TCGA–KO–8409 (KICH) unknown
TCGA–QG–A5Z2 (COAD) unknown

Figure S19



MSI-H
MSI-H (P)
MSI-L/MSS
Relative frequency (%)
30
25
20
15
10
5
0
ACN>A
CCN>A
GCN>A
TCN>A
ACN>G
CCN>G
GCN>G
TCN>G
ACN>T
CCN>T
GCN>T
TCN>T
ATN>A
CTN>A
GTN>A
TTN>A
ATN>C
CTN>C
GTN>C
TTN>C
ATN>G
CTN>G
GTN>G
TTN>G