Comprehensive Analysis of Indels in Whole-genome Microsatellite Regions and Microsatellite Instability across 21 Cancer Types

Akihiro Fujimoto[1,2], Masashi Fujita[1], Takanori Hasegawa[3], Nagahide Matsubara[4], Kazuhiro Maejima[1], Kaoru Nakano[1], Aya Oku-Sasaki[1], Jing Hao Wong[2], Yuichi Shiraishi[5], Satoru Miyano[3,5], Seiya Imoto[3], Naohiro Tomita[4], Michael R Stratton[6], Steven G Rosen[7], and Hidewaki Nakagawa[1], on behalf of the PCAWG Signature Working Group and The PCAWG Network

[1]*Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Science, Japan*
[2]*Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, Japan*
[3]*Health Intelligence Center, Institute of Medical Sciences, The University of Tokyo, Japan*
[4]*Division of Lower Gastrointestinal Surgery, Department of Surgery, Hyogo College of Medicine, Nishinomiya, Japan.*
[5]*Human Genome Center, Institute of Medical Sciences, The University of Tokyo, Japan*
[6]*Welcome Trust Sanger Institute, UK*
[7]*Center for Computational Biology, Duke-NUS Medical School, Singapore*

[*]*Correspondence should be addressed to Akihiro Fujimoto (akihiro.fujimoto@riken.jp) and Hidewaki Nakagawa (hidewaki@ims.u-tokyo.ac.jp).*

Abstract

Microsatellites are repeats of 1-6bp units and ~10 million microsatellites have been identified across the human genome. Microsatellites are sensitive to DNA mismatch errors, and therefore have been used to detect cancers with mismatch repair deficiency. To reveal the mutational landscape of the microsatellite repeat regions at the genome level, we analyzed approximately 9.2 million microsatellites in 2,717 whole genomes of pan-cancer samples across 21 tissue types. First, we developed an insertion and deletion caller that takes into consideration the error patterns of different types of microsatellites. Among the 2,717 pan-cancer samples, our analysis identified 31 samples, including colorectal, uterus, and stomach cancers, with higher microsatellite mutation rate (≥ 0.03), which we defined as microsatellite instability (MSI) cancers. Next, we found 20 highly-mutated microsatellites to detect MSI cancers with high sensitivity. Validation of the marker set in an independent cancer cohort showed that the marker set was effective. Third, we found that replication timing and DNA shape were significantly associated with mutation rates of the microsatellites. Analysis of germline variation of the microsatellites suggested that the amount of germline variations and somatic mutation rates were correlated. Lastly, analysis of mutations in the mismatch repair genes showed that somatic SNVs and short indels had larger functional impact than germline mutations and structural variations. Our analysis provides a comprehensive picture of mutations in the microsatellite regions, and reveals possible causes of mutations, as well as provides a useful marker set for MSI detection.

Introduction

Recent large-scale whole genome sequencing studies have revealed the complexity of the mutational landscape of the cancer genome (1-4). In cancer genomes, various types of mutations, such as SNVs (single nucleotide variants), short indels (insertions and deletions), genomic rearrangements, copy number alternations, insertion of retrotransposons, and virus genome integrations, have been identified, and their oncogenic roles have been characterized (1-5). Additionally, genome sequencing studies have revealed the molecular basis of somatic mutations (6-9). However, somatic mutations in microsatellites or repeat sequences have not been well characterized in a large whole genome sequencing cohort due to difficulties in detecting mutations in them accurately by present short-read sequencing technologies.

A microsatellite is defined as a tract of repetitive DNA motif composed of short repeating unit (10). The mutation rate of microsatellites has been known to be higher than other genomic regions because of DNA polymerase slippage during DNA replication and repair (10). Due to their fragility, microsatellites are used as markers of genomic instability in cancer (11). In cancer genetics studies, microsatellite instability (MSI) has been used for molecular diagnosis of Lynch syndrome and cancers with mismatch repair deficiency (11). Furthermore, MSI-positive tumors are generally burdened with higher numbers of somatic mutations and present many mutation-associated neoantigens, which might be recognized by the immune system. Presently, MSI can also be used as a marker to predict the effect of immune therapy (12). The MSI phenotype is most common in colorectal cancers, stomach cancers and uterine endometrial cancers (10-15%), although it has also been observed across many tumor types at a few % (11). The MSI phenotype is defined by the presence of somatic indels of the 2-5 microsatellite makers, whereby BAT25/26, mononucleotide microsatellites are widely used to establish MSI status (11).

Regardless of the clinical importance of microsatellite, large-scale analysis of somatic changes in microsatellites across various type of cancers is limited for whole genome sequencing (WGS) data (13, 14). To reveal the whole genome mutational landscape of microsatellite regions, we here analyzed indels in microsatellites for 2,913 ICGC pan-cancer samples from 21 tissues (15). We developed a method to detect somatic indels in microsatellite regions, selected an appropriate parameter for our purpose, and identified indels in microsatellite regions. We identified MSI-positive

samples and factors affecting the mutation rate of microsatellites, as well as highly-mutated microsatellites. We also analyzed the association of mutation rate of microsatellite with somatic and germline mutations in DNA repair genes, and compared mutational signatures between MSI and other samples. Our analysis provides a comprehensive picture of mutations in the microsatellite regions, and reveals possible causes of mutations, as well as provides a useful marker set for MSI detection.

Results

Identification of microsatellite regions in the genome

We detected microsatellites using three methods (MsDetector, Tandem Repeat Finder, and MISA software) (16-18). To exclude microsatellites potentially arising from read mapping errors, we selected microsatellites based on the uniqueness of flanking sequences and pattern of repeat. A total of 9,292,677 microsatellites were used for subsequent analyses. In the selected microsatellites, MISA software identified a larger number of microsatellites than other methods (**Supplementary Fig. 1**).

Error rate estimation of microsatellites

During library preparation and sequencing processes, indel errors can be introduced in reads containing short repeats by PCR, due to replication slippage of DNA polymerases. Since the error rates should depend on the length and type of microsatellites, we first estimated error rates of different types of microsatellites. Using sequence data of chrX from 32 normal tissues of male individuals, we estimated the error rate among the different types and lengths of microsatellites. As the male chrX is hemizygotic, we can infer the error rate without the influence of heterozygous polymorphisms (19, 20). As expected, error rates depended on the unit and length of the microsatellites, and longer microsatellites had higher error rates (**Supplementary Fig. 2**). In all types of the microsatellites, deletion errors were more frequent than insertion errors, and smaller changes of unit number were predominant (**Supplementary Fig. 2**). These results suggest that PCR or sequencing processes are prone to induce short deletion errors. Error rates of microsatellites between 5-9bp in length were very low (< 0.2%), while those of longer microsatellites were higher (> 5%). Two bp repeat had higher error rates than other microsatellites (**Supplementary Fig. 2**). The A/T type of microsatellite was observed to have higher error rates compared to G/C microsatellites (**Supplementary**

Fig. 2). Since the estimated error rates were quite different among the types and the lengths of microsatellites, we used the difference between error rates to detect somatic indels in the microsatellites. We generated a table of error rates for analyzing mutations in microsatellite regions based on the estimated error rates (**Supplementary Table 2**).

Validation with simulation data sets and setting of thresholds

Mutations in microsatellite regions were identified based on likelihoods (see Materials and Methods). To estimate false positive and false negative rates, and to select appropriate parameters, we generated simulation data sets. We used sequence reads mapped on chrX of male individuals. First, we determined the genotype of each microsatellite on chrX. Since chrX is hemizygotic in male, we considered the dominant reads as the true genotype of each sample. We then mixed chrX reads from two male individuals, and identified variations in the microsatellite regions with our algorithm. By comparing the true genotypes and genotypes from the mixed data, we estimated the false positive and negative rates. The false positive and negative rates were varied according to the likelihood values (L), and higher L had higher false negative and lower false positive rates (**Supplementary Fig. 3a, b, c**). Longer microsatellites tended to have a higher false negative rate, as longer microsatellites tend to be covered by a smaller number of reads (**Supplementary Fig. 3d**). Based on the analysis, L was set to -8 for cancer samples and -1 for matched normal samples.

Analysis of indels in the microsatellite regions in pan-cancer samples

We analyzed the whole genome sequence data of 2,917 pan-cancer samples (21) with our method, and compared them against somatic and germline variants detected by the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (22). To compare our results with somatic consensus indels from the four PCAWG indel callings, we gathered indels located ±5bp in the microsatellite regions in the PCAWG calls. On average, 1826.5 indels were detected by our indel caller in the microsatellite regions. Of these, 1185.1 were found only by our indel caller (**Supplementary Fig. 4a**), suggesting higher sensitivity of our indel calls compared to the other PCAWG callers for microsatellite regions. We then compared the number of indels in the microsatellite regions between our indel caller and PCAWG callers. In the microsatellite regions, the number of indels uniquely identified by our indel caller was significantly correlated with that of

commonly identified indels (identified by ≥ two PCAWG callers) (Pearson's product moment correlation coefficient; r = 0.90, *p-value* < 10^{-16}) (**Supplementary Fig. 4b**). We further performed experimental validation with Japanese liver cancer samples for the randomly selected microsatellite mutation candidates by capillary electrophoresis. The false discovery rate of our method was estimated to be 7% (27/29) (**Supplementary Fig. 5**). These results indicate that our microsatellite indel caller can identify indels in the microsatellite region effectively.

Microsatellites covered by ≥ 15 reads in ≥ 2,500 samples (7,650,128 microsatellites) were subjected to further analysis, and samples with ≥ 6,000,000 testable microsatellites were used for the analysis (2,717 samples) (**Supplementary Table 3**). On average, 7,407,000 microsatellites were analyzed in each sample. We compared the proportion of mutated samples for each microsatellite. Most of the microsatellites in whole genome were not mutated in the pan-cancer samples (**Fig. 1a**), we therefore selected 198,578 microsatellites with mutated samples proportions ≥ 0.001 (more than 2-3 mutated samples in the pan-cancer samples) and considered them as informative microsatellites. The proportions of mutated samples were different among the types and lengths of the microsatellite, with A/T microsatellites more frequently mutated than other types of microsatellites (**Fig. 1b**). The proportions of mutated microsatellites were significantly lower in exonic and intronic regions, but higher in non-genic regions (**Supplementary Fig. 6**), suggesting that indels in microsatellites located in genic region could be repaired by transcription-coupled repair mechanism.

We next defined samples with proportions of mutated microsatellites ≥ 0.03 as microsatellite instability (MSI) samples (n=31) (**Fig. 1c**), and others as microsatellite stable (MSS) samples. As expected, colorectal (CR), uterus (UT), and stomach (ST) cancers had a larger number of MSI samples, but MSI was also observed in a minority of samples in liver (LI), pancreas (PA), ovary (OV), kidney (KI), and skin (SK) cancers (**Fig. 1c**). The mutation pattern of microsatellites was different between the MSI and MSS samples. In the MSI samples, deletions were more predominant compared to insertions (**Fig. 1d and 1e**). We compared the proportion of mutated microsatellites, the number of somatic SNVs and somatic indels (**Fig. 1f and 1g**). Although the numbers of SNVs were not strongly correlated with the proportion of mutated microsatellites (r = 0.19), the number of somatic indels were clearly correlated (r = 0.97) (**Fig. 1f and 1g**), suggesting that microsatellite and non-microsatellite indels are affected by common

mechanisms of mutation and repair, and that the analysis of microsatellites can predict samples with large number of non-microsatellite indels.

Mutability of microsatellite

Recent studies have suggested that epigenetic factors, DNA structures, and sequence motif influence the mutation rate (8, 9, 23, 24). However, little is known about factors that influence the mutability of microsatellites. We first analyzed the replication timings and microsatellite mutation rates (**Fig. 2a-d**). The late replicating regions had lower mutation rates for insertions and deletions in the MSS samples and insertions in the MSI samples (**Fig. 2abc**). However, an opposite pattern was observed for deletions of the MSI samples; the early replicating regions had higher mutation rates (**Fig. 2d**).

For a more detailed analysis, we performed a multiple regression. In our analysis, the majority of mutated microsatellites were A/T mononucleotide repeat as previously reported, suggesting that the fragility is primarily determined by the base composition (20) (**Fig. 1b**). Therefore, to find other factors that associate with the mutation rate of microsatellites, we selected 1,000 highly mutated A/T microsatellites of 10-30bp length in the reference genome, and analyzed them for deletions and insertions in the MSS and MSI samples. We considered replication timing, nuclear lamina binding region, G-quadruplexes, and predicted DNA shapes. Nuclear lamina binding regions are known to be associated with genomic fragile sites (23), while G-quadruplexes can cause replication errors (24). The impact of DNA shapes was not well known, but one DNA shape parameter (ORChID2) had been reported to be associated with mutation rate of somatic indels (9). Multiple regression analyses for these factors showed that the length of microsatellite, replication timing, and several DNA shapes were significantly associated with the mutation rate of microsatellites (**Supplementary Table 4**).

The predicted DNA shapes of the flanking sequences were significantly associated with the proportion of mutated samples (**Fig 2e-h and Supplementary Table 4**). Several DNA shape features such as ORChID2 (OH Radical Cleavage Intensity), HelT (Helix Twist), Opening, MGW (Minor Groove Width), Rise, ProT (Propeller Twist), Roll, and Slide, were significantly associated with the prevalence of insertions and deletions in microsatellite regions (**Fig 2e-h, Supplementary Fig. 7, Supplementary Table 4**). The nuclear lamina binding region and G-quadruplexes were not significantly associated (**Supplementary Table 4**). The adjusted R^2 values of the

multiple regression analysis were 0.25 in the deletions of MSI, 0.14 in the insertions of MSI, 0.28 in the deletions of MSS and 0.29 in the insertions of MSS (**Supplementary Table 4**).

Microsatellites are highly polymorphic and have been used as genetic makers for population genetics studies (10). To evaluate the genetic polymorphism, we estimated the heterozygosity of each microsatellite locus in normal tissues. The proportion of mutated samples in cancers and the heterozygosity in normal tissues was significantly correlated (Pearson's product moment correlation coefficient; r = 0.31, *P-value* < 10^{-16}) (**Fig 2i**), which indicates that genetic variations and somatic mutations are influenced by same factors.

Highly mutated microsatellites

We compared mutability of each microsatellite between the MSI and MSS samples, and selected top 20 highly mutated microsatellites with the highest mutation rate (**Supplementary Table 5**). We performed a clustering analysis with these microsatellite markers, and confirmed that they perfectly distinguished the MSI and MSS samples in the CR, UT and ST cancers (**Fig 3a-c**).

For validation in an independent cohort, we collected ~300 fresh frozen colorectal cancer tissues at Hyogo Medical University Hospital and performed MSI test using the most established microsatellite markers (BAT25 and BAT26) at the clinical level. Among them, we detected 26 MSI-positive cases which were used for validation of our highly mutated microsatellite. For 26 MSI and 10 MSS colorectal cancers, we analyzed microsatellite mutations of 12 microsatellite markers among the 20 highly mutated microsatellites and recurrently mutated coding microsatellites (see below). **Figure 3d** shows the presence/absence of the indels in each of these highly mutated microsatellites. Analysis of recurrently mutated coding microsatellites revealed that several microsatellites were enriched in MSI samples (**Fig. 4**). Microsatellites in *ASTE1, TGFBR2,* and *KIAA2018* were mutated in more than 80% of the validation samples (*ASTE1* 23/26*, TGFBR2* 20/26*,* and *KIAA2018* 21/26). These results indicated that these microsatellite regions were as sensitive as BAT25/26, and had a technical advantage over known MSI markers in availability in whole genome sequencing.

Genes with large number of mutated microsatellites

To find genes with high mutation rates, we tested the total number of indels in microsatellites for each gene across the 21 tumor types. We counted the number of mutated microsatellites and the total number of analyzed microsatellites for each gene, and identified genes with larger numbers of mutated microsatellites than others. After adjusting for multiple testing, 1,134 genes had significantly larger numbers of mutated microsatellites for at least one tissue (*q-value* < 0.01) (**Supplementary Table 6**). Of these genes, *ALB,* which is known to be highly expressed in liver, showed the larger number of mutated microsatellite (Fisher's exact test; *q-value* = 6.5×10^{-15}, odds ratio = 65.1) in liver cancer (LI) (**Supplementary Fig. 8**). A previous study suggested that some cell linage-specific highly-expressed genes, including *ALB* in liver, had recurrent short indels (25). This result is consistent with the precious study, and strong DNA damage in cell linage-specific highly-expressed genes would influence mutation rate of microsatellites (**Supplementary Table 6**).

Recurrently mutated microsatellites in the coding region

We compared recurrently mutated microsatellites in the coding regions between MSS and MSI samples (**Fig 4, Supplementary Table 7**). Microsatellites in *ACVR2A* and *TGFBR2*, which have been reported to be frequently mutated in MSI tumors (13, 14, 20), were recurrently mutated in 60% and 47% of the MSI samples, respectively. In addition, microsatellites in *ASTE1*, *KIAA2018*, *LIN1* and *CDH26* were also mutated in more than 50% of the MSI samples. Mutations in microsatellites in Cosmic Cancer Genes (*MSH6, JAK1, BLM, IL7R,* and *CSF3R*) were identified as MSI specific mutations. Of these, indels in *MSH6,* which is a mismatch repair gene, is likely to cause the MSI phenotype (11). Mutations in *JAK1* in MSI cancers were reported to associate with tumor immune evasion (26). In the MSS samples, microsatellites in *APC,* and *TCF12* was mutated only in MSS samples, suggesting that these mutations cause cancer without genomic instability.

Although many of the recurrently mutated coding microsatellites have been reported by whole genome or exome sequencing studies (**Supplementary Table 7**) (13, 14, 20, 27), our analysis identified new genes with recurrently mutated microsatellites. Of these, the *GINS1* gene encodes a subunit of DNA replication complex (28). *MBD4* has been reported to contribute to tumorigenesis and work as a modifier of MMR-deficient cancer (29). *BLM* is included in the Cosmic Cancer Gene database and

has functions in DNA replication and DNA double-strand break repair (30). These results suggest that mutations in microsatellites can work as driver events.

We also validated these microsatellites or homopolymer mutations in our colorectal cancer samples (26 MSI and 10 MSS samples) by Sanger sequencing, and found that these coding mutations occurred specifically in MSI tumors as well as the new microsatellite makers in non-coding regions described above (**Fig. 3b**)

Mutation in mismatch repair (MMR) and proofreading genes

We analyzed the association between the proportion of mutated microsatellites and somatic and germline variants of eight DNA repair genes (*MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6*, *PMS2*, *POLE* and *POLD1*). First, we focused on stop gain, splice site, nonsense mutations, and gene-disrupting structural variations in tumor and matched normal samples (**Fig 5a-d**). Two samples in CR and UT had a loss of function germline mutations in the *MSH2* or *PSM2* gene, suggesting that Lynch syndrome should cause cancers in these patients (**Fig 5a and 5c**) (11). The number of samples with somatic SNVs and indels in these genes were significantly enriched in the MSI samples (Fisher's exact test CR; *P-value* = 2.0×10^{-8}, Odds ratio = 10.6, ST; *P-value* = 1.2×10^{-4}, Odds ratio = 49.0, UT; *P-value* = 1.4×10^{-5}, Odds ratio = 40.6), while those with germline variants and structural variations (SVs) were not significantly enriched. These results suggest that most of the MSI phenotypes in cancer were mainly caused by somatic short indels or somatic SNVs. Germline SVs of *PMS2* were frequently observed in MSS tumors, indicating less impact of *PMS2* to DNA mismatch repair deficiency or MSI (31).

Most MSI samples had larger numbers of somatic SNVs (**Fig. 1c**), because of functional deficiency of MMR genes. However, 40 MSS samples had larger numbers of somatic SNVs than the average number of SNVs in the MSI samples (151,816.6 SNVs). Of these, 8 had somatic missense mutation in the exonuclease domain of *POLE* (residues 268-471) (32) (**Supplementary Fig. 9, Supplementary Table 8**), suggesting that exonuclease domain mutations of *POLE* were associated with a large number of SNVs in MSS, not MSI.

Association with somatic mutational signatures in PCAWG

We compared mutational signatures found in single base substitution (SBS), doublet base substitution (DBS) as well as insertions and deletions (ID) between the MSI and MSS samples (**Fig 6a-c, Supplementary Fig. 10, Supplementary Table 9**). The PACWG signature analysis detected 49 SBS, 11 DBS, and 17 ID signatures (33). We compared the fraction of each mutational signature between MSI and MSS samples in CR (Colon/Rectum), ST (Stomach), and UT (Uterus), and found that six SBS signatures (SBS5, SBS15, SBS20, SBS21, SBS26 and SBS44), one ID signature (ID2), and four DBS signatures (DBS3, DBS7, DBS8 and DBS10) were significantly different among the MSI and MSS samples in at least one cancer type (Wilcoxon signed-rank test, *q-value* < 0.05). Most of these mutational signatures have been reported to be associated with tumors with defective DNA mismatch repair (33), expect for DBS3 and DBS8. This time we found that DBS3 and DBS8 were associated with MSI. DBS3 was also associated with the mutations in exonuclease domain of *POLE* in the current study (Wilcoxon signed-rank test, *q-value* < 0.05) (**Supplementary Table 10**), and no etiology has been proposed for DBS8, which was observed in ES (Esophagus adenocarcinoma) and CR (33). In the ID signatures, the fraction of ID2 (A/T deletion) was significantly different between MSI and MSS in CR, ST, and UT, which is consistent with an excess of A/T indels in the microsatellite regions (**Fig.1b**).

Neo-antigen load from the microsatellite or repeat coding regions

MSI cancer is known to show specific immune reactions such as Crohn-like reaction and diffused infiltration of lymphocytes in pathology (34), and PD-1 inhibiting immune therapy is a highly effective treatment for all types of tumors showing MSI (12). Its specific immune-reaction should be related with neo-antigen burden, and we thus calculated the neo-antigen burdens of MSI and MSS tumors by using somatic mutations and HLA genotypes, taking into account neo-peptides generated from indel or frameshift mutations of the coding microsatellites. The number of predicted neo-antigens in MSI tumors (median, 393) was significantly higher than the MSS tumors (median, 11; *P-value* = 5.9×10^{-20}) (**Supplementary Fig. 11**), which is consistent with the mutational burden. Interestingly, while 95% of neo-antigens were derived from SNVs in MSS tumors, in MSI tumors 51% of the predicted neo-antigens were derived from short indels and 5% were derived from short indels of the microsatellites.

Discussion

Due to the clinical importance of MSI phenotypes, exome sequencing and small-scaled WGS studies were performed for the MSI samples (13, 14, 20, 27). These studies identified recurrently mutated microsatellites and driver genes mainly located in coding regions, as well as created an algorithm to find MSI with smaller number of microsatellite sets. However, the detailed analysis for factors that influence on mutation rate and validation of the selected microsatellite marker sets in independent cohorts was limited. Here, we performed an analysis of the microsatellite mutations in the largest whole genome sequencing cohort with the largest number of microsatellites to date to characterize microsatellite mutations and MSI tumors at the genome-wide level.

To identify microsatellite regions in human genome, we used results from three software programs (Tandem repeat finder, MS detector, and MISA) (16-18). After filtering, we obtained 9,292,677 microsatellites in genome for analysis. Among the methods for detecting microsatellites, MISA software identified the largest number of microsatellites (17). Although most of them were short and may not be considered as microsatellites by the other two methods, they contained highly mutated repeat regions. It has been reported that the rate of mutation of longer microsatellites is higher than that of shorter microsatellites (35). However, the current short-reads sequencing technologies are unable to analyze longer microsatellites. We therefore decided to prioritize shorter repeats as the main targets for the current WGS study. Furthermore, regarding the most established mononucleotide microsatellite markers for MSI such as Bethesda markers, sequence read numbers mapped to these complicated regions were quite small and we could not analyze their mutation in the most of the WGS (**Supplementary Fig. 12**). Alternatively, in our study we detected highly-mutated short microsatellites (**Supplementary Table 5**) and they could be useful to clinical diagnosis of MSI with the current short read technologies.

The analysis of WGS and the validation study in an independent cohort found 20 novel microsatellite markers (**Fig. 3a-d**), which can be used to predict tumors with high mutational burden (**Fig. 3a and 3b**). However, the mutation rate of microsatellite was highly correlated with the total number of indels, but not strongly correlated with the total number of SNVs (**Fig. 1cfg**), indicating that samples with larger numbers of SNVs can be found in MSS samples. As the analysis of neo-antigens showed that SNVs

can also produce a larger number of neo-antigens (**Supplementary Fig. 11**), identification of these samples is also important for diagnosis. In our analysis, the analysis of MSS samples with a larger number of SNVs showed that mutations in the exonuclease domain of *POLE* can partly explain the high mutation rate of SNVs, not indels and microsatellites (**Supplementary Fig. 9**). Therefore, analysis of mutations in the *POLE* gene in MSS samples can identify more tumors with high mutational burdens (36).

Our WGS analysis found high rates of short deletions in the microsatellites of both MSI and MSS samples (**Fig. 1de**). The excess of deletion events was also observed in our validation study with MSI colon cancers (**Fig. 3 and Supplementary Fig. 13**). Therefore, the excess of short deletions should not be due to a bias of our mutation calling method, and can be considered as a common feature of cancers. A microsatellite mutation model suggests that deletions are generated by a misalignment loop in the template strand, and insertions subsequently generated in the nascent strand (10). During the DNA replication of cancer cells, template strands could exist as single strands for a longer period compared to nascent strands, resulting in a higher chance to generate misaligned loop structures, which would induce larger numbers of deletions.

The analysis of replication timing showed a different pattern between deletions in MSI and others (**Fig. 2a-d, Supplementary Table 4**). As observed in the SNVs, replication timing and mutation rate were positively corelated with the insertions and deletions of MSS samples and insertions of MSI samples (**Fig. 2a-d**). It is suggested that early replicating regions are more accessible for DNA repair machineries, resulting in more chances for repair (8). However, deletions in microsatellites were enriched in the early replicating regions of MSI samples **(Fig. 2d)**. A recent exome sequencing study also reported the inverse correlation between the microsatellite indels and replication timing in MSI tumors (20). Since MSI tumors should have defects in their DNA mismatch repair machinery, this result would reflect the pattern of mutation without DNA mismatch repair. In early replication, template strands may exist as single strands for a longer period, facilitating the occurrence of deletions.

In addition to replication timing, DNA shape parameters were also associated with the mutation rate of microsatellites. Microsatellites with lower HelT (Helix twist), higher ProT (Propeller twist), higher Roll, higher Slide had higher insertion rates (**Fig. 2e-h, Supplementary Fig. 7**). Microsatellites with higher HelT, higher Opening, higher

ProT, higher Roll, lower Slide had higher deletion rates (**Fig. 2e-h, Supplementary Fig. 7**). Since DNA shapes were associated in both the MSI and the MSS tumors, they would affect the fragility of DNA strand, and mainly influence the mutation generation instead of the repair process.

In the present study, we considered that the mutation rate of microsatellites is mainly influenced by the following; fragility of DNA sequences (length and unit type of microsatellite, and DNA shape) (**Fig. 1b**, **Fig. 2e-h**), activity of DNA repair machinery (mutations or activities in mismatch repair machinery genes) (**Fig. 5)**, DNA damage against cell linage-specific highly-expressed genes (**Supplementary Fig. 8, Supplementary Table 6**), and accessibility of DNA repair machinery (DNA replication timing) (**Fig. 2a-d**). Furthermore, since the mutation rate in cancer was correlated with the heterozygosity of germline variations, these factors would also affect the mutation rate in germline variations (**Fig. 2i**).

The current study analyzed somatic indels in microsatellite regions in the largest WGS cohort to date. We found a microsatellite marker set to detect MSI, factors that influence the mutation rate of microsatellite, genes with recurrently mutated microsatellite, and the influence of somatic mutations in MMR and proof-reading genes have on MSI. Our analysis provides a mutational landscape of microsatellites in cancer samples for future clinical applications.

Acknowledgements

The super-computing resource "SHIROKANE" was provided by Human Genome Center, The University of Tokyo (http://sc.hgc.jp/shirokane.html). This work was supported partially by RIKEN President's Fund 2011, Grand-in-aid for RIKEN CGM and IMS, Grant-in-Aid for Scientific Research on Innovative Areas from JSPS grants (25134717, 25670375, 23114001, 15H04814), Project for Cancer Research and Therapeutic Evolution (P-CREATE) (Grant Number 16cm0106519h0001, to H, N.). and Platform Program for Promotion of Genome Medicine (Grant Number 18km0405207h0003, to A.F.) in the Japan Agency for Medical Research and Development (AMED).

URLs.

Software for somatic mutation; MIMcall (https://github.com/afujimoto/MIMcall)

Software for germline variation; MIVcall (https://github.com/afujimoto/MIVcall)
DNA shape parameters; http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/
Lamina binding region;
https://www.nature.com/article-assets/npg/nature/journal/v453/n7197/extref/nature06947-s2.txt
Software for estimating location of G-quadruplex;
https://github.com/dariober/bioinformatics-cafe/blob/master/fastaRegexFinder.py
Replication timing;
https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwRepliSeq
Cosmic cancer genes; https://cancer.sanger.ac.uk/cosmic

Author Contributions

Study design: A. F., S.G. R., M.R.S., and H. N. Data analysis: A. F., M. F., H. T., S. I., Y. S., S. M. and H. N. Molecular analysis: M. K., N. K. and H. N. Sample acquisition: M. N. and T. N. Manuscript writing: A. F., M. F. and H. N.

Materials and Methods

Samples and data

Whole genome sequencing data was obtained by the International Cancer Genome Consortium (ICGC) pan-cancer project (15). The list of analyzed samples is shown in the **Supplementary Table 1 and 3**. Datasets of somatic point mutations, short indels, structural variants (SVs), and copy number alterations were generated as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (21, 22). Overall, 2,834 samples with whole genome data are represented in the PCAWG datasets, spanning a range of cancer types (bladder, sarcoma, breast, liver-biliary, cervix, leukemia, colorectal, lymphoma, prostate, esophagus, stomach, central nervous system, head/neck, kidney, lung, melanoma, ovary, pancreas, thyroid, and uterus). The consensus somatic SNVs and short indels in PCAWG samples were determined using different algorithms; calls made by at least two algorithms were used in downstream analyses (22).

Definition of microsatellite region for the analysis

We determined microsatellite regions using MsDetector, Tandem Repeat Finder and MISA software (16-18). Microsatellite regions defined by the Tandem Repeat Finder were obtained from the UCSC database (16). Identification of microsatellite with MISA was done by (unit size) = 1 to 5, (minimum number of repeats) = 5 and (max difference between 2 microsatellites) = 10. Since these three methods used different algorithms to define microsatellite, we first defined the repeat unit of each microsatellite. We divided each region with different lengths (1-6bps), and calculated the entropy of the character string. The length with the lowest entropy was selected as the unit length of each microsatellite region. For the analysis of the microsatellite, we filtered microsatellite regions according to the following criteria; (1) the proportion of the most frequent unit ≥ 0.8, (2) distance between closest neighboring microsatellite ≥ 30bp, (3) if the microsatellite regions were detected by 2 or more methods, we selected the longest one and discarded others, and (4) upstream and downstream flaking sequences (100bp) of each microsatellite were mapped against human reference genome (GRCh37) by blat software (37) with the options of -stepSize=5 and -repMatch=2253, and microsatellites that had ≥ 90bp of flanking sequences mapped to different positions were removed. As a result of this selection procedure, 8,817,054 autosomal microsatellites remained and were used for the subsequent analyses (**Supplementary Fig. 1**).

Error rate estimation of each repeat unit

To identify somatic indels in the microsatellite regions, we first estimated the error rate of the different types of repeat units. For this purpose, we used data from chromosome X of 32 male normal samples, because chrX is homozygotic and error rates can be estimated without the influence of heterozygous polymorphisms (19, 20). The estimated error rates are shown in **Supplementary Fig. 2**.

Identification of change of repeat unit from whole genome sequence

Microsatellites are repeat sequences and mapping errors can influence the accuracy of detection. To remove possible mapping errors, we removed improper pairs and reads with low mapping quality (< 30), as well as reads with large (> 550bp) or small (< 100bp) read pair distance.

We counted the number of repeat units in each microsatellite region. We then determined the genotype of the matched normal tissues and detected somatic indels by comparing the genotype of the normal and cancer samples. To distinguish the mutation or variation from sequencing errors, we incorporated the binomial distribution with the estimated error rates (**Supplementary Fig. 2**) and calculated a likelihood for each variant candidate.

For normal samples, we calculated the likelihood for the second most frequent number of repeat (**Supplementary Fig. 2**).

$$L_i = \binom{n}{r}(p_i)^r(1-p_i)^{n-r},$$

where n is the total number of reads that cover the microsatellite, r is the number of reads containing ith repeat, and p_i is the estimated error rate of the ith repeat. If the likelihood is lower than a threshold value, the genotype was assumed to be heterozygous for the major repeat and second major repeat. We next calculated the likelihood for the number of repeat in the cancer. If the likelihood of the non-germline repeat was lower than a threshold value, we defined the repeat as a somatic indel candidate.

To find the appropriate likelihood threshold values, we applied this algorithm on data from other male chrX. Based on the estimated false positive and negative rates, we set -1 and -8 for germline genotyping and somatic mutation calling (**Supplementary Fig. 3**) respectively. Based on the comparison, we set L=-8 for tumor, and L=-1 for

matched normal samples. We selected the microsatellites that were covered by ≥ 15 reads in both the cancer and the matched normal samples. Additionally, we selected somatic indels with variant allele frequency in cancer ≥ 0.15 and number of support reads in cancer ≥ 2 and ≤ 1 in the normal samples.

Estimation of false discovery rate

We randomly selected 29 somatic MS mutations detected in RK004, RK249 and RK308 and performed validation with a previously reported method (36). Amplicons were analyzed using the ABI PRISM ® 3100 Genetic Analyzer (Applied Biosystems), and GeneMapper software (Applied Biosystems). Validation for the selected microsatellites was also performed using the Sanger sequencing method (**Supplementary Fig. 5**).

Selection of highly mutated MSs

To select microsatellites, we compared the number of mutated samples in microsatellite instability (MSI) and microsatellite stable (MSS) samples for each microsatellite. The numbers were compared using Fisher's exact test. Nine microsatellites with odds ratio ((number of mutated samples in the MSI)/ (number of un-mutated samples in the MSI)/ (number of mutated samples in the MSS)/ (number of un-mutated samples in the MSS)) ≥ 500 and proportion of the mutated samples in the MSI samples ≥ 0.8 were selected and genotyped in the additional samples. Three microsatellites were selected from highly-mutated microsatellites in the MSI samples. Ten microsatellites were selected from recurrently mutated coding microsatellites.

Validation of the colon cancer cohort

For new microsatellite makers or mutated regions, we designed primers to amplify them and performed PCR on DNA from 36 frozen colorectal cancer tissues and the corresponding normal colon tissues (resected at Hyogo Medical University Hospital). Ethical committees at RIKEN, the Hyogo Medical University Hospital and all groups participating in this study approved this work. The analysis by the most established microsatellite makers for MSI, BAT25 and BAT26, identified 26 MSI samples and 10 MSS samples. The amplicons were sequenced by ABI 3100 capillary sequencers or MiSeq after illumina adaptor ligation.

Comparison of mutational signatures between the MSS and MSI

Mutational signatures and their proportions were obtained from the result of the PCAWAG signature working group (31). The proportion of each signature was compared between MSI and MSS samples in CR, UT, and ST with the Wilcoxon signed-rank test. Multiple testing adjustment was done using Benjamini and Hochberg's FDR method (39).

Identification of genes with recurrently mutated microsatellites

To identify highly-mutated genes, we compared the mutation rate of microsatellites in each gene. We compared the total number of analyzed microsatellite and total number of mutated microsatellites in intron and exon in each gene. We also counted the total number of analyzed microsatellite and total number of mutated microsatellites in the entire genome in MSS samples for each cancer type; (total number of mutated microsatellite in gene *i* in all MSS samples in cancer *j*)/(total number of unmutated microsatellite in gene *i* in all MSS samples in cancer *j*) and (total number of mutated microsatellite in entire genome in all MSS samples in cancer *j*)/(total number of unmutated microsatellite in entire genome *i* in all MSS samples in cancer *j*) were compared with Fisher's exact test. In this analysis, we could obtain genes with larger numbers of mutated microsatellites compared to the entire genome. Multiple testing adjustment was done using Benjamini and Hochberg's FDR method (39).

Prediction of neoantigens

HLA genotyping from WGS data were generated as part of the PCAWG project. Somatic point mutations, non-MS indels detected by the PCAWG project, and MS indels detected by our method were combined and annotated using ANNOVAR. Mutant peptides of length 8–11 residues were assessed for their binding affinity (IC_{50}) to the HLA class I of matched patients using NetMHCpan-3.0 (40). Mutant peptides of $IC_{50} <$ 50 nM were predicted as neoantigens.

FIGURE LEGENDS

Figure1. Pattern of somatic indels in microsatellite regions.
(a) Mutation rate of each microsatellite. 7,650,128 microsatellites were sorted by the proportion of mutated samples. The red box indicates informative microsatellites defined in this study (proportion of mutated samples ≥ 0.001). **(b)** Mutation rate among different microsatellites. **(c)** Comparison of mutation rate of microsatellites, number of somatic SNVs and indels in different types of cancer. MSI samples are shown in red. **(d-e)** Pattern of insertions (positive change in repeat length in x-axis) and deletions (negative change in repeat length in x-axis) in microsatellites between the MSI **(d)** and MSS samples **(e)**. Correlation between the mutation rate of microsatellites and the number of somatic SNVs. Pearson's product-moment correlation; $r = 0.19$, *p-value* = 1.6×10^{-23}. MSI samples are shown in red. **(f)** Correlation between the mutation rate of microsatellites and the number of somatic indels. Pearson's product-moment correlation; $r = 0.97$, *p-value* < 1.0×10^{-200}. MSI samples are shown in red.

Figure 2. Analysis of mutation rate of each microsatellite.
(a-d) Association of replication timing with insertion and deletion rates. The edges of the boxes represent the 25th and 75th percentile values. The whiskers represent the most extreme data points, which are no more than 1.5 times the interquartile range from the boxes. **(a)** Insertions in MSS samples (χ^2 test; *p-value* < 1×10^{-200}), **(b)** Insertions in MSI samples (χ^2 test; *p-value* < 1×10^{-200}), **(c)** Deletions in MSS samples (χ^2 test; *p-value* < 1×10^{-200}), **(d)** Deletions in MSI samples (χ^2 test; *p-value* < 1×10^{-200}). **(e-h)** Association of DNA shape with mutation rate. Top 1000 A/T microsatellites with 10-30bp in length were used for the analysis. The microsatellites were divided into the three categories based on the mutation rate. **(e)** HelT (Helix twist) of insertions in MSS samples. **(f)** ORChID2 (the ·OH Radical Cleavage Intensity) of insertions in MSS samples, **(g)** ProT (Propeller Twist) of deletions in MSI samples, **(h)** Slide of deletions in MSI samples.
(i) Germline variation and somatic mutation rates of microsatellites. The amount of germline variation was estimated using heterozygosity of normal tissues. X-axis; the proportion of mutated samples in each microsatellite (somatic mutation rate). Y-axis; heterozygosity of each microsatellite. Both rates were positively correlated (Pearson's product-moment correlation; $r = 0.31$, *p-value* < 1.0×10^{-100}).

Figure 3. Highly-mutated microsatellite markers. Result of clustering analysis with the top 20 microsatellites. **(a)** CR (Colon/Rectum) cancer, **(b)** ST (Stomach) cancer, **(c)** UT (Uterus) caner. **(d)** Result of validation study in the independent colon cancer cohort (n=36). MSI was defined by BAT25 or BAT26 (upper panel). Twelve new microsatellite markers (middle panel) were successfully amplified and analyzed for indel mutations by MiSeq (**Supplementary Fig. 12**). Fourteen highly-mutated microsatellites or homopolymers in coding regions were analyzed (lower panel). These mutations occurred specifically in MSI colorectal cancers. Mutation status are shown in brown (mutated) and white (WT; wild type) in the panel. Number of mutated samples in the MSI samples are shown in bar plot (right).

Figure 4. Proportion of mutated samples in coding microsatellites. x-axis; proportion of mutated samples in the MSI samples. y-axis; proportion of mutated samples in the MSS samples. Cosmic cancer genes are shown in red.

Figure 5. Mutation in mismatch repair (MMR) and proofreading genes. Mutated genes, and mutation rate of microsatellites are shown.

Figure 6. Comparison of mutational signatures between the MSI and MSS samples. **(a)** CR (Colon/Rectum) cancer, **(b)** ST (Stomach) cancer, **(c)** UT (Uterus) caner. Signatures showing significant difference between the MSI and MSS samples are shown in rectangles in the legends (Wilcoxon signed-rank test, *q-value* < 0.05).

References

1. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjaerde OC, Langerod A, Ringner M, Ahn SM, Boyault S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GK, Jang SJ, Jones DR, Kim HY, King TA, Krishnamurthy S, Lee HJ, Lee JY, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporte I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodriguez-Gonzalez FG, Romieu G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, Van den Eynden GG, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, van't Veer L, Tutt A, Knappskog S, Tan BK, Jonkers J, Borg A, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver MJ, Martens JW, Borresen-Dale AL, Richardson AL, Kong G, Thomas G, Stratton MR (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534: 47-54
2. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, Patch AM, Kakavand H, Alexandrov LB, Burke H, Jakrot V, Kazakoff S, Holmes O, Leonard C, Sabarinathan R, Mularoni L, Wood S, Xu Q, Waddell N, Tembe V, Pupo GM, De Paoli-Iseppi R, Vilain RE, Shang P, Lau LMS, Dagg RA, Schramm SJ, Pritchard A, Dutton-Regester K, Newell F, Fitzgerald A, Shang CA, Grimmond SM, Pickett HA, Yang JY, Stretch JR, Behren A, Kefford RF, Hersey P, Long GV, Cebon J, Shackleton M, Spillane AJ, Saw RPM, Lopez-Bigas N, Pearson JV, Thompson JF, Scolyer RA, Mann GJ (2017) Whole-genome landscapes of major melanoma subtypes. Nature 545: 175-180
3. Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, Grobner S, Segura-Wang M, Zichner T, Rudneva VA, Warnatz HJ, Sidiropoulos N, Phillips AH, Schumacher S, Kleinheinz K, Waszak SM, Erkek S, Jones DTW, Worst BC, Kool M, Zapatka M, Jager N, Chavez L, Hutter B, Bieg M, Paramasivam N, Heinold M, Gu Z, Ishaque N, Jager-Schmidt C, Imbusch CD, Jugold A, Hubschmann D, Risch T, Amstislavskiy V, Gonzalez FGR, Weber UD, Wolf S, Robinson GW, Zhou X, Wu G, Finkelstein D, Liu Y, Cavalli FMG, Luu B,

Ramaswamy V, Wu X, Koster J, Ryzhova M, Cho YJ, Pomeroy SL, Herold-Mende C, Schuhmann M, Ebinger M, Liau LM, Mora J, McLendon RE, Jabado N, Kumabe T, Chuah E, Ma Y, Moore RA, Mungall AJ, Mungall KL, Thiessen N, Tse K, Wong T, Jones SJM, Witt O, Milde T, Von Deimling A, Capper D, Korshunov A, Yaspo ML, Kriwacki R, Gajjar A, Zhang J, Beroukhim R, Fraenkel E, Korbel JO, Brors B, Schlesner M, Eils R, Marra MA, Pfister SM, Taylor MD, Lichter P (2017) The whole-genome landscape of medulloblastoma subtypes. Nature 547: 311-317

4. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, Gotoh K, Ariizumi S, Wardell CP, Hayami S, Nakamura T, Aikata H, Arihiro K, Boroevich KA, Abe T, Nakano K, Maejima K, Sasaki-Oku A, Ohsawa A, Shibuya T, Nakamura H, Hama N, Hosoda F, Arai Y, Ohashi S, Urushidate T, Nagae G, Yamamoto S, Ueda H, Tatsuno K, Ojima H, Hiraoka N, Okusaka T, Kubo M, Marubashi S, Yamada T, Hirano S, Yamamoto M, Ohdan H, Shimada K, Ishikawa O, Yamaue H, Chayama K, Miyano S, Aburatani H, Shibata T, Nakagawa H (2016) Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 48: 500-9
5. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, Abril G, Pereira G, Makohon-Moore A, Looijenga LH, Gillis AJ, Hruban RH, Anders RA, Romans KE, Pandey A, Iacobuzio-Donahue CA, Vogelstein B, Kinzler KW, Kazazian HH, Jr., Solyom S (2015) Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res 25: 1536-45
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson

M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR (2013) Signatures of mutational processes in human cancer. Nature 500: 415-21

7. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, Campbell PJ, Vineis P, Phillips DH, Stratton MR (2016) Mutational signatures associated with tobacco smoking in human cancer. Science 354: 618-622
8. Tubbs A, Nussenzweig A (2017) Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell 168: 644-656
9. Chen W, Zhang L (2015) The pattern of DNA cleavage intensity around indels. Sci Rep 5: 8333
10. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5: 435-45
11. Boland CR, Goel A (2010) Microsatellite instability in colorectal cancer. Gastroenterology 138: 2073-2087 e3
12. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, Wong F, Azad NS, Rucki AA, Laheru D, Donehower R, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Greten TF, Duffy AG, Ciombor KK, Eyring AD, Lam BH, Joe A, Kang SP, Holdhoff M, Danilova L, Cope L, Meyer C, Zhou S, Goldberg RM, Armstrong DK, Bever KM, Fader AN, Taube J, Housseau F, Spetzler D, Xiao N, Pardoll DM, Papadopoulos N, Kinzler KW, Eshleman JR, Vogelstein B, Anders RA, Diaz LA, Jr. (2017) Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 357: 409-413
13. Kim TM, Laird PW, Park PJ (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 155: 858-68
14. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ (2017) A molecular portrait of microsatellite instability across multiple cancers. Nat Commun 8: 15180
15. International Cancer Genome Consortium. (2010) International network of cancer genome projects. Nature 464: 993-8
16. Girgis HZ, Sheetlin SL (2013) MsDetector: toward a standard computational tool for DNA microsatellites detection. Nucleic Acids Res 41: e22
17. Hause RJ, Pritchard CC, Shendure J, Salipante SJ (2016) Classification and

characterization of microsatellite instability across 18 cancer types. Nat Med 22: 1342-1350
18. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-80
19. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet 42: 931-6
20. Maruvka YE, Mouw KW, Karlic R, Parasuraman P, Kamburov A, Polak P, Haradhvala NJ, Hess JM, Rheinbay E, Brody Y, Koren A, Braunstein LZ, D'Andrea A, Lawrence MS, Bass A, Bernards A, Michor F, Getz G (2017) Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat Biotechnol 35: 951-959
21. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD (2018) Pan-cancer analysis of whole genomes. bioRxiv
22. Goldman et al. (2018) Online resources for PCAWG data exploration, visualization, and discovery. bioRxiv
23. Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD (2012) A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Res 22: 993-1005
24. Lemmens B, van Schendel R, Tijsterman M (2015) Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers. Nat Commun 6: 8909
25. Imielinski M, Guo G, Meyerson M (2017) Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. Cell 168: 460-472 e14
26. Albacker LA, Wu J, Smith P, Warmuth M, Stephens PJ, Zhu P, Yu L, Chmielecki J (2017) Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion. PLoS One 12: e0176181
27. Kondelin J, Gylfe AE, Lundgren S, Tanskanen T, Hamberg J, Aavikko M, Palin K, Ristolainen H, Katainen R, Kaasinen E, Taipale M, Taipale J, Renkonen-Sinisalo L, Jarvinen H, Bohm J, Mecklin JP, Vahteristo P, Tuupanen S, Aaltonen LA, Pitkanen E (2017) Comprehensive Evaluation of Protein Coding Mononucleotide

Microsatellites in Microsatellite-Unstable Colorectal Cancer. Cancer Res 77: 4078-4088

28. Ueno M, Itoh M, Kong L, Sugihara K, Asano M, Takakura N (2005) PSF1 is essential for early embryogenesis in mice. Mol Cell Biol 25: 10528-32
29. Tricarico R, Cortellino S, Riccio A, Jagmohan-Changur S, Van der Klift H, Wijnen J, Turner D, Ventura A, Rovella V, Percesepe A, Lucci-Cordisco E, Radice P, Bertario L, Pedroni M, Ponz de Leon M, Mancuso P, Devarajan K, Cai KQ, Klein-Szanto AJ, Neri G, Moller P, Viel A, Genuardi M, Fodde R, Bellacosa A (2015) Involvement of MBD4 inactivation in mismatch repair-deficient tumorigenesis. Oncotarget 6: 42892-904
30. Patel DS, Misenko SM, Her J, Bunting SF (2017) BLM helicase regulates DNA repair by counteracting RAD51 loading at DNA double-strand break sites. J Cell Biol 216: 3521-3534
31. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, Dowty JG, White DM, Hopper JL, Baglietto L, Jenkins MA, de la Chapelle A (2008) The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. Gastroenterology 135: 419-28
32. Church DN, Briggs SE, Palles C, Domingo E, Kearsey SJ, Grimes JM, Gorman M, Martin L, Howarth KM, Hodgson SV, Collaborators N, Kaur K, Taylor J, Tomlinson IP (2013) DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. Hum Mol Genet 22: 2820-8
33. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG, Stratton MR (2018) The Repertoire of Mutational Signatures in Human Cancer. bioRxiv
34. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, Fishel R, Lindor NM, Burgart LJ, Hamelin R, Hamilton SR, Hiatt RA, Jass J, Lindblom A, Lynch HT, Peltomaki P, Ramsey SD, Rodriguez-Bigas MA, Vasen HF, Hawk ET, Barrett JC, Freedman AN, Srivastava S (2004) Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst 96: 261-8

35. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44: 1161-5
36. Mehnert JM, Panda A, Zhong H, Hirshfield K, Damare S, Lane K, Sokol L, Stein MN, Rodriguez-Rodriquez L, Kaufman HL, Ali S, Ross JS, Pavlick DC, Bhanot G, White EP, DiPaola RS, Lovell A, Cheng J, Ganesan S (2016) Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. J Clin Invest 126: 2334-40
37. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.
38. Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18: 233-4
39. Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300 (1995).
40. Nielsen M, Andreatta M (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med 8: 33

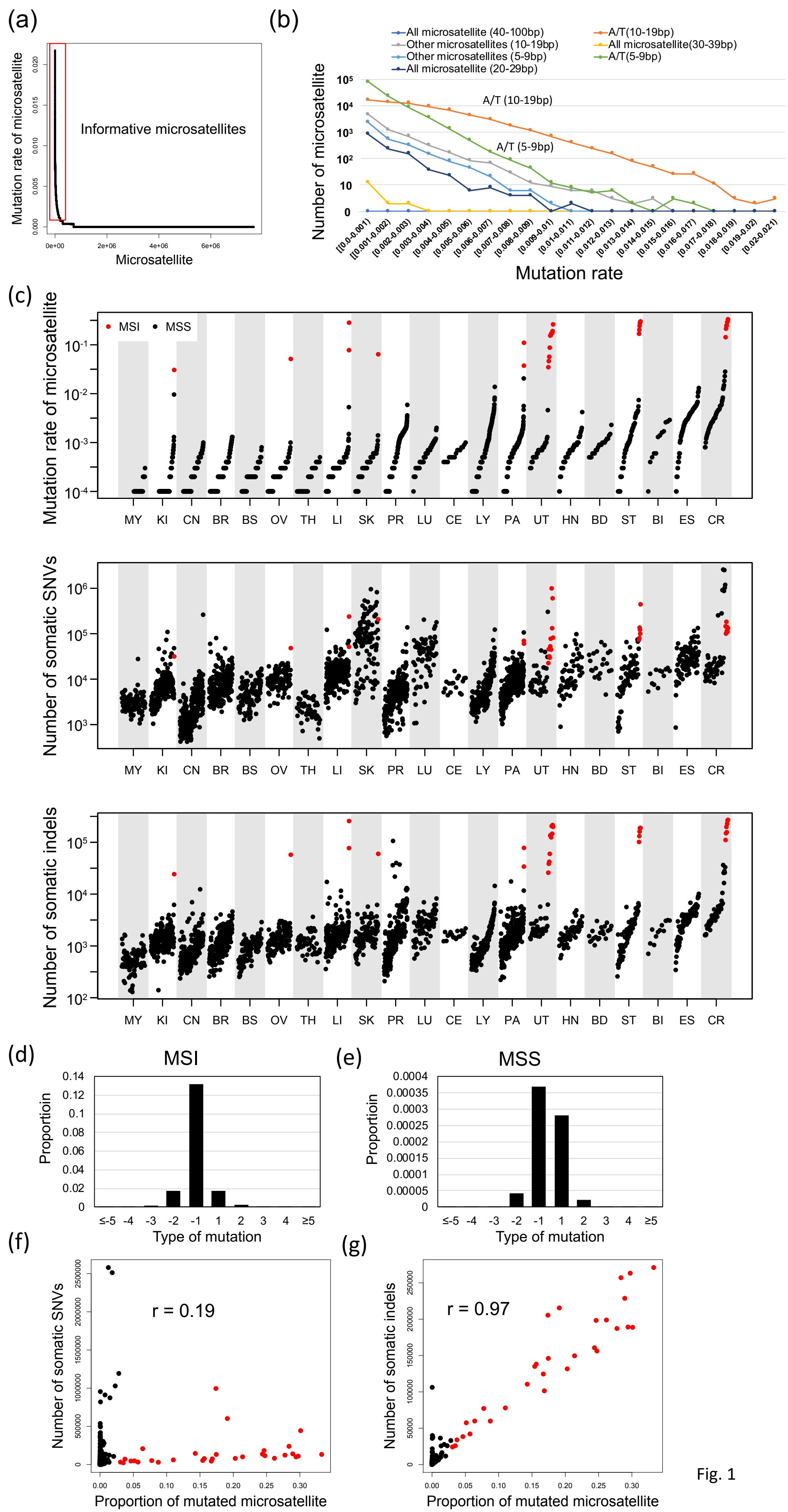

(a)
Mutation rate of microsatellite
Informative microsatellites
Microsatellite
(b)
All microsatellite (40-100bp)
Other microsatellites (10-19bp)
Other microsatellites (5-9bp)
All microsatellite (20-29bp)
A/T(10-19bp)
All microsatellite(30-39bp)
A/T(5-9bp)
A/T (10-19bp)
A/T (5-9bp)
Number of microsatellite
Mutation rate
(c)
MSI
MSS
Mutation rate of microsatellite
Number of somatic SNVs
Number of somatic indels
MY KI CN BR BS OV TH LI SK PR LU CE LY PA UT HN BD ST BI ES CR
(d)
MSI
Proportioin
Type of mutation
(e)
MSS
Proportioin
Type of mutation
(f)
r = 0.19
Number of somatic SNVs
Proportion of mutated microsatellite
(g)
r = 0.97
Number of somatic indels
Proportion of mutated microsatellite

Fig. 1

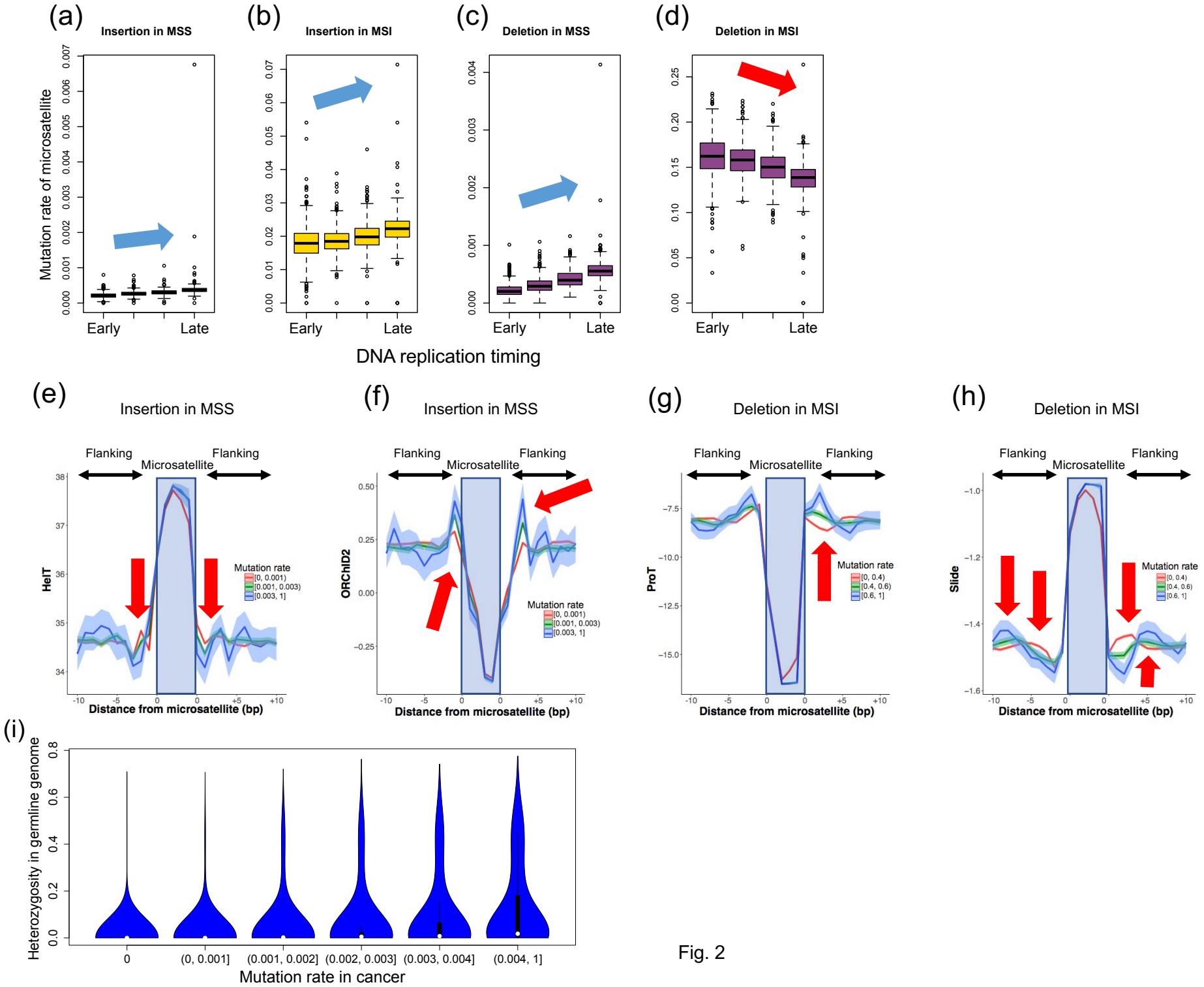
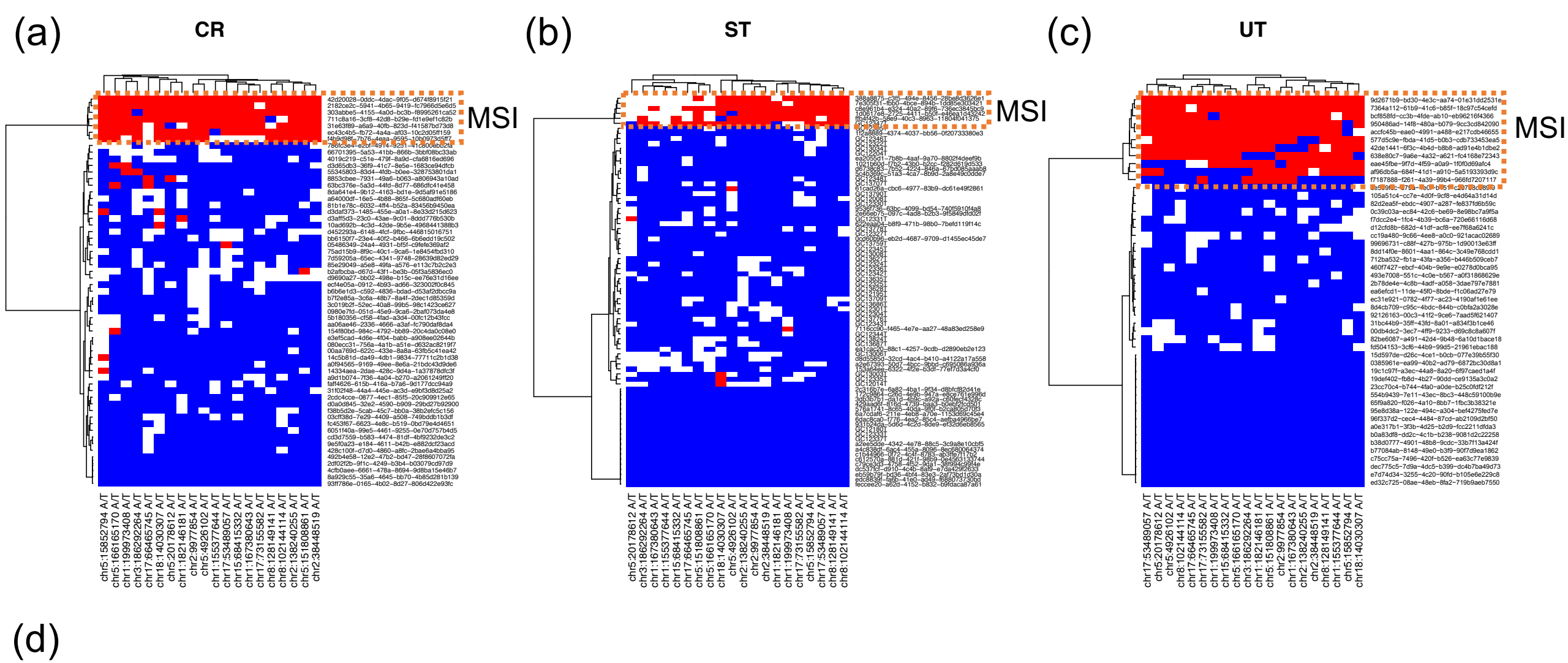

(a)
Insertion in MSS
(b)
Insertion in MSI
(c)
Deletion in MSS
(d)
Deletion in MSI
Mutation rate of microsatellite
Early
Late
DNA replication timing
(e)
Insertion in MSS
(f)
Insertion in MSS
(g)
Deletion in MSI
(h)
Deletion in MSI
Flanking
Microsatellite
Flanking
HelT
ORChID2
ProT
Slide
Mutation rate
[0, 0.001)
[0.001, 0.003)
[0.003, 1]
[0, 0.4)
[0.4, 0.6)
[0.6, 1]
Distance from microsatellite (bp)
(i)
Heterozygosity in germline genome
0
(0, 0.001]
(0.001, 0.002]
(0.002, 0.003]
(0.003, 0.004]
(0.004, 1]
Mutation rate in cancer

Fig. 2


(a)
CR
MSI
(b)
ST
MSI
(c)
UT
MSI

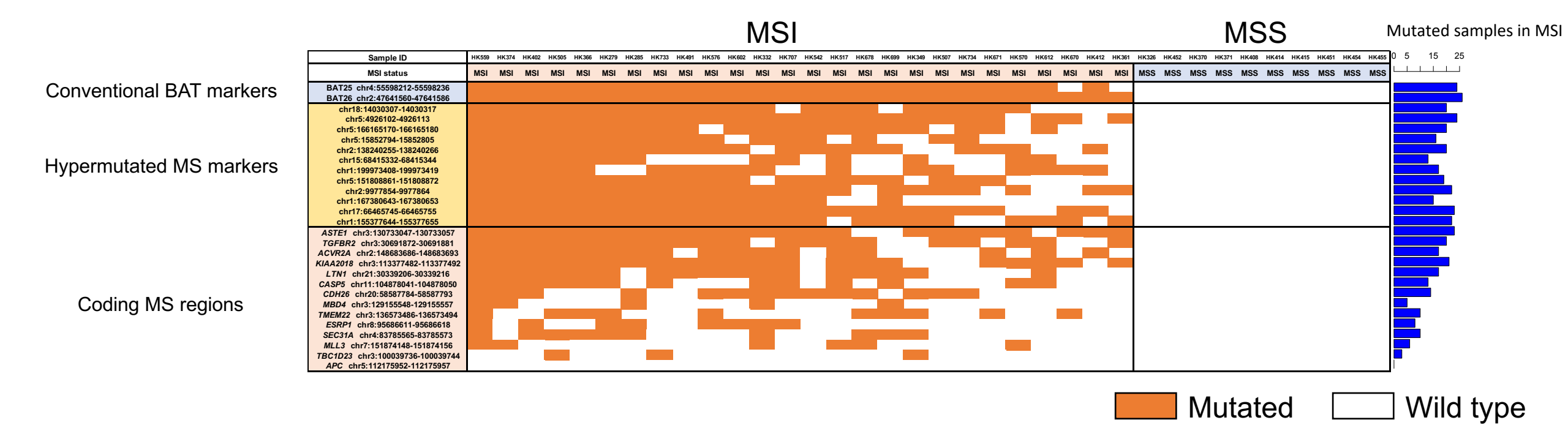

(d)
MSI
MSS
Mutated samples in MSI
Sample ID
MSI status
Conventional BAT markers
BAT25 chr4:55598212-55598236
BAT26 chr2:47641560-47641586
Hypermutated MS markers
chr18:14030307-14030317
chr5:4926102-4926113
chr5:166165170-166165180
chr5:15852794-15852805
chr2:138240255-138240266
chr15:68415332-68415344
chr1:199973408-199973419
chr5:151808861-151808872
chr2:9977854-9977864
chr1:167380643-167380653
chr17:66465745-66465755
chr1:155377644-155377655
Coding MS regions
ASTE1 chr3:130733047-130733057
TGFBR2 chr3:30691872-30691881
ACVR2A chr2:148683686-148683693
KIAA2018 chr3:113377482-113377492
LTN1 chr21:30339206-30339216
CASP5 chr11:104878041-104878050
CDH26 chr20:58587784-58587793
MBD4 chr3:129155548-129155557
TMEM22 chr3:136573486-136573494
ESRP1 chr8:95686611-95686618
SEC31A chr4:83785565-83785573
MLL3 chr7:151874148-151874156
TBC1D23 chr3:100039736-100039744
APC chr5:112175952-112175957
Mutated
Wild type

Fig. 3

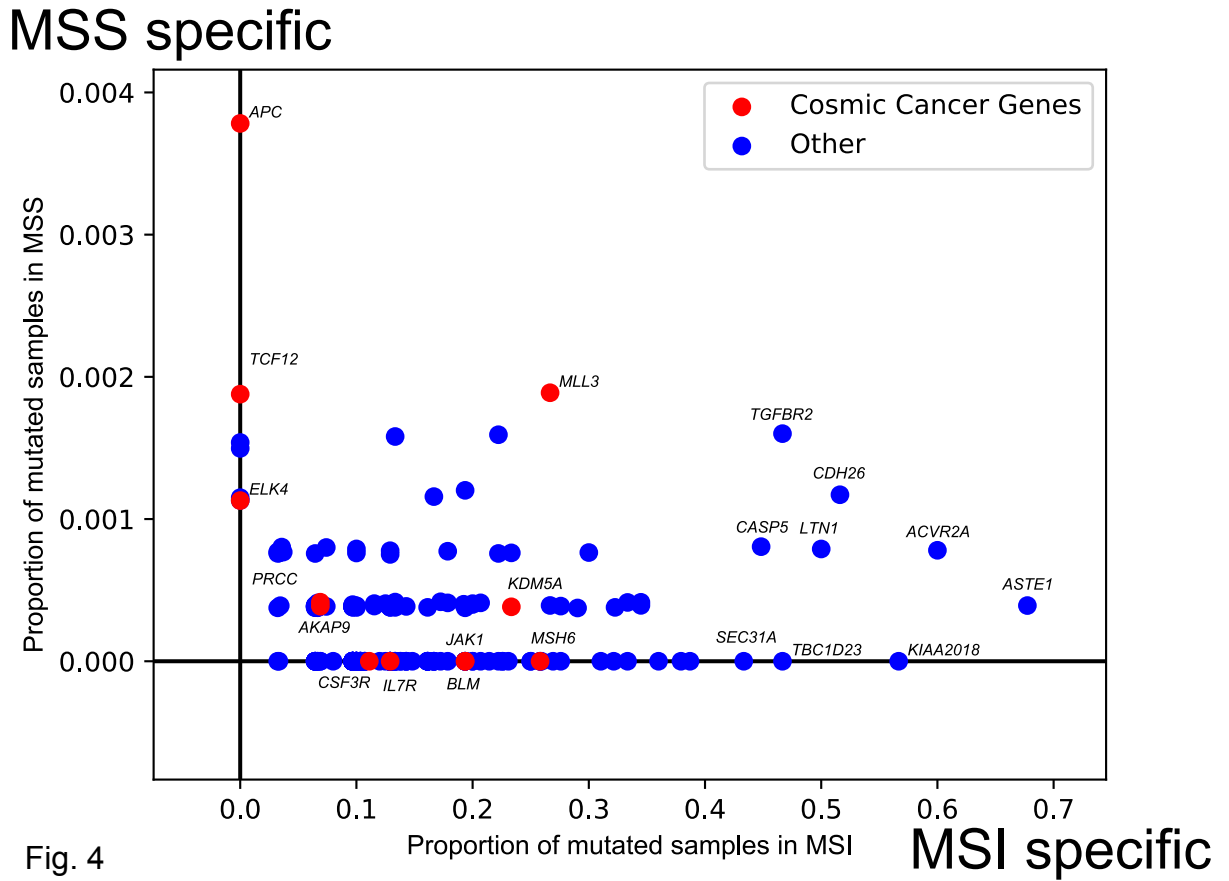

MSS specific
Cosmic Cancer Genes
Other
Proportion of mutated samples in MSS
Proportion of mutated samples in MSI
MSI specific
0.004
0.003
0.002
0.001
0.000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
APC
TCF12
ELK4
MLL3
TGFBR2
CDH26
CASP5
LTN1
ACVR2A
ASTE1
PRCC
KDM5A
AKAP9
JAK1
MSH6
SEC31A
TBC1D23
KIAA2018
CSF3R
IL7R
BLM

Fig. 4

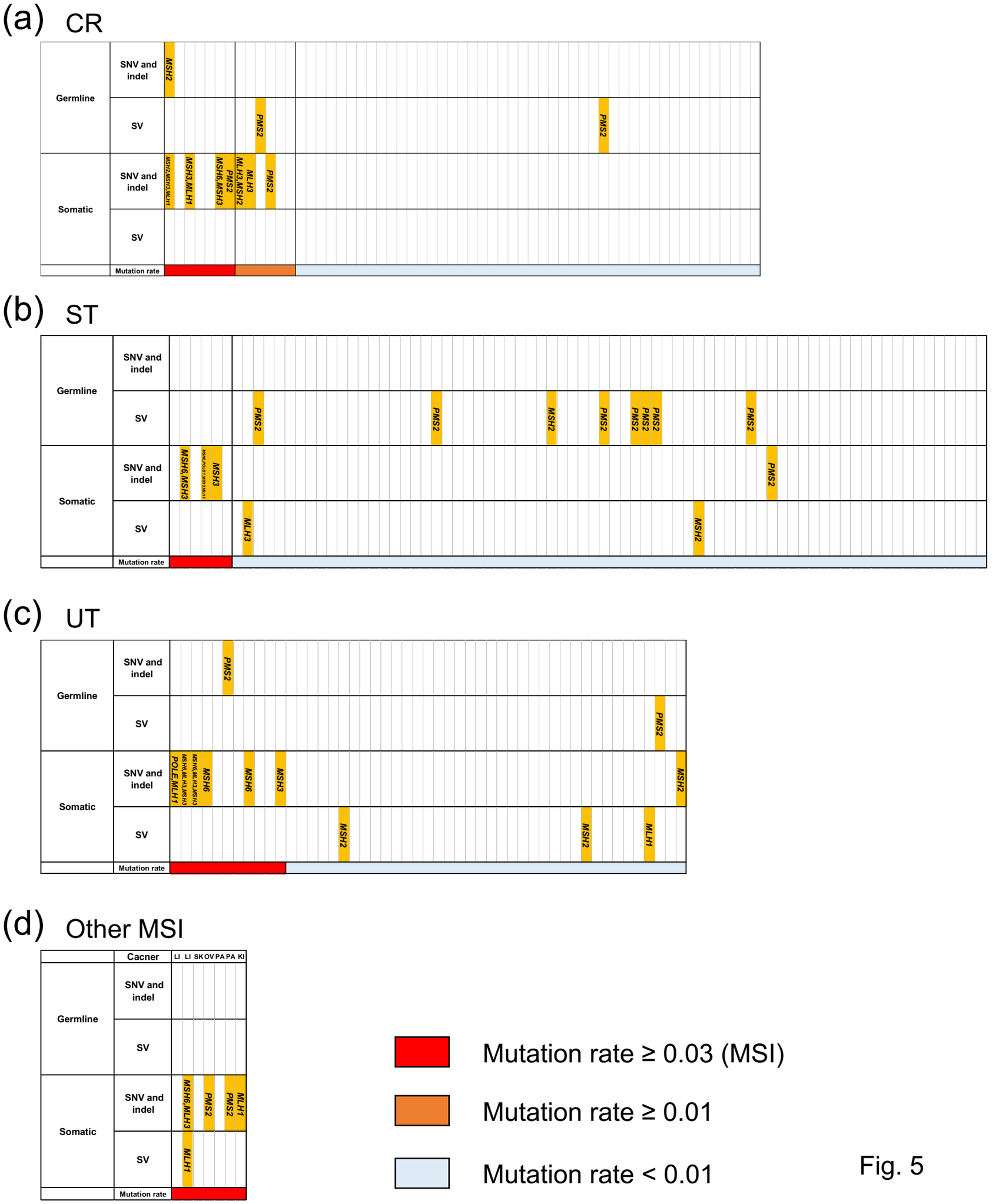

(a) CR
Germline
SNV and indel
MSH2
SV
PMS2
PMS2
Somatic
SNV and indel
MSH2,MSH3,MLH1
MSH3,MLH1
MSH6,MSH3
PMS2
MLH3,MSH2
MLH3
PMS2
SV
Mutation rate
(b) ST
Germline
SNV and indel
SV
PMS2
PMS2
MSH2
PMS2
PMS2
PMS2
PMS2
PMS2
Somatic
SNV and indel
MSH6,MSH3
MSH3
MSH6,POLD1,MSH3,MLH1
PMS2
SV
MLH3
MSH2
Mutation rate
(c) UT
Germline
SNV and indel
PMS2
SV
PMS2
Somatic
SNV and indel
POLE,MLH1
MSH6,MLH3,MSH3
MSH6,MLH3,MSH3
MSH6
MSH6
MSH3
MSH2
SV
MSH2
MSH2
MLH1
Mutation rate
(d) Other MSI
Cacner
LI LI SK OV PA PA KI
Germline
SNV and indel
SV
Somatic
SNV and indel
MSH6,MLH3
PMS2
PMS2
MLH1
SV
MLH1
Mutation rate
Mutation rate ≥ 0.03 (MSI)
Mutation rate ≥ 0.01
Mutation rate < 0.01

Fig. 5

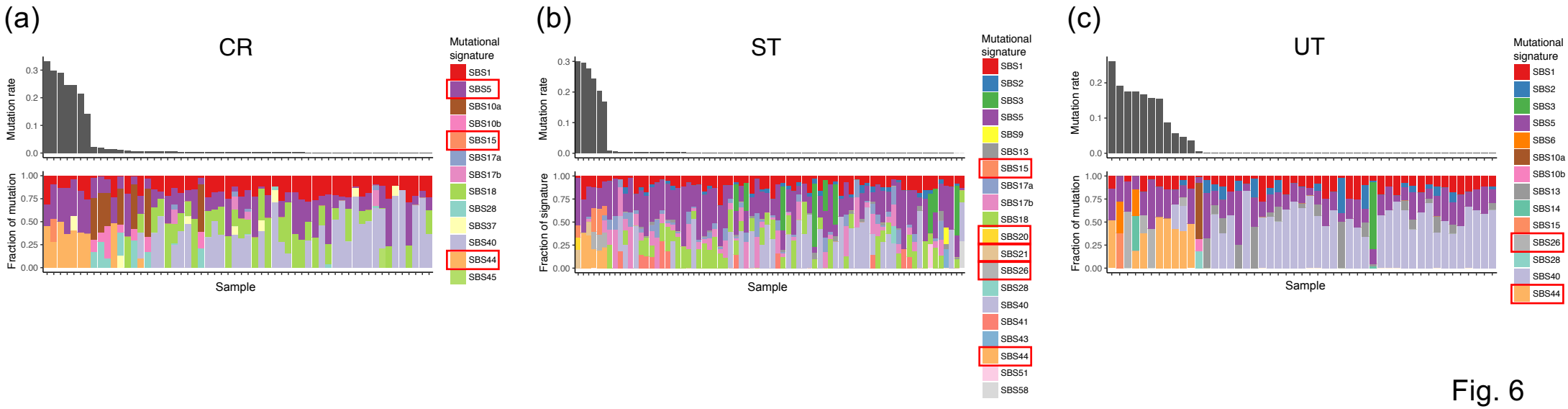

(a)
CR
Mutation rate
Fraction of mutation
Sample
Mutational signature
SBS1
SBS5
SBS10a
SBS10b
SBS15
SBS17a
SBS17b
SBS18
SBS28
SBS37
SBS40
SBS44
SBS45
(b)
ST
Mutation rate
Fraction of signature
Sample
Mutational signature
SBS1
SBS2
SBS3
SBS5
SBS9
SBS13
SBS15
SBS17a
SBS17b
SBS18
SBS20
SBS21
SBS26
SBS28
SBS40
SBS41
SBS43
SBS44
SBS51
SBS58
(c)
UT
Mutation rate
Fraction of mutation
Sample
Mutational signature
SBS1
SBS2
SBS3
SBS5
SBS6
SBS10a
SBS10b
SBS13
SBS14
SBS15
SBS26
SBS28
SBS40
SBS44

Fig. 6