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], Jing Hao Wong[2], Kazuhiro Maejima[1], Aya Oku-Sasaki[1], Kaoru Nakano[1], Yuichi Shiraishi[5], Satoru Miyano[3,4], Seiya Imoto[3], Michael R Stratton[5], Steven G Rosen[6], and Hidewaki Nakagawa[1], and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes 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]*Human Genome Center, Institute of Medical Sciences, The University of Tokyo, Japan*
[5]*Welcome Trust Sanger Institute, UK*
[6]*Center for Computational Biology, Duke-NUS Medical School, Singapore*

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

Abstract

Microsatellites are repeats of 1-6bp units and ~10 million microsatellites have been identified across the human genome. Microsatellites are vulnerable to DNA mismatch errors, and have thus 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 20.1 billion microsatellites in 2,717 whole genomes of pan-cancer samples across 21 tissue types. Firstly, we developed a new insertion and deletion caller (MIMcall) 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 in genome-wide level. Next, we found 20 highly-mutated microsatellites that can be used to detect MSI cancers with high sensitivity. 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 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 alterations, 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 accurately detecting mutations using presently available short-read sequencing technologies.

A microsatellite is defined as a tract of repetitive DNA motif composed of short repeating units (10). The mutation rate of microsatellites has been known to be higher than other genomic regions due to 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 neo-antigens, 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).

Irrespective 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). In the current study, we analyzed indels in microsatellites for 2,913 ICGC pan-cancer samples from 21 tissues (15) to reveal the whole genome mutational landscape of microsatellite regions. 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 microsatellites 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 repeats. A total of 9,292,677 microsatellites were used for subsequent analyses. Within these selected microsatellites, it was observed that the MISA software identified a larger number compared to other methods (**Supplementary Fig. 1**).

Error rate estimation of microsatellites

During library preparation and sequencing processes, indel errors can be introduced by PCR in reads containing short repeats, due to replication slippage of DNA polymerases. Since the error rates should depend on the length and type of microsatellites, we first estimated the error rates of different types of microsatellites. The type of microsatellites was defined by length of the microsatellite region in the reference genome and repeat unit (see **Materials and Methods**). Using sequence data of chrX from 32 normal tissues of male individuals, we estimated the error rate among different types and lengths of microsatellites. As the male chrX is hemizygotic, the error rate can be inferred without the influence of heterozygous polymorphisms (19, 20). As expected, error rates depended on the unit and length of the microsatellites, with longer microsatellites having higher error rates (**Supplementary Fig. 2**). In all types of 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 within the reference genome were very low (< 0.2%), while those of longer microsatellites were higher (> 5% error rate for 20-100bp of microsatellites in the reference genome length). Two bp repeats 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 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 by using 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**). To identify somatic mutations in microsatellites, we require reads that completely cover target microsatellites. The length of reads is about 100bp, therefore, longer microsatellites have fewer reads covering them compared to shorter microsatellites, and thus have a lower sensitivity (**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 (MIMcall) in the microsatellite regions. Of these, 1185.1 were found only by MIMcall (**Supplementary Fig. 4a**), suggesting a higher sensitivity of our indel calls compared to the other PCAWG callers for microsatellite regions.

PCAWG calls removed repetitive regions to achieve highly accurate mutation calling, therefore our result can complement the PCAWG calls. 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; $r = 0.90$, $p\text{-}value < 10^{-16}$) (**Supplementary Fig. 4b**). We further performed experimental validation with Japanese liver cancer samples for the mutation candidates in longer microsatellites by capillary electrophoresis (**Supplementary Fig. 5**). The false discovery rate of our method was estimated to be 7% (2/29) (**Supplementary Fig. 5**). These results indicate that MIMcall 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 genomes were not mutated in the pan-cancer samples (**Fig. 1a**), we therefore selected 198,578 microsatellites with proportions of mutated samples ≥ 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 microsatellite, with A/T microsatellites more frequently mutated than other types (**Fig. 1b**). The proportions of mutated microsatellites were significantly lower in exonic and intronic regions, but higher in non-genic regions (**Supplementary Fig. 6**). Microsatellites in CDS (coding sequence) regions would evolve to be more stable to avoid mutations and this would cause lower mutation rates in the CDS regions. Lower mutation rates in intronic regions suggests the influence of transcription-coupled repair.

We next defined samples with proportions of mutated microsatellites ≥ 0.03 as microsatellite instability (MSI) samples (n=31) in genome-wide level (**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 for liver (LI), pancreas (PA), ovary (OV), kidney (KI), and skin (SK) cancers (**Fig. 1c**). The proportions of MSI samples were 11.9% for colorectal (7/59, 95% C.I. 4.9%-22.9%), 7.7% for stomach (6/78, 95% C.I. 2.8%-16.0%)

and 22.0% for uterine (11/50, 95% C.I. 11.5%-36.0%) cancers, which are consistent with previous studies (23-25). We additionally analyzed Bethesda markers, which is a conventional marker set for MSI definition. However, the number of sequence reads that mapped to these regions was quite small and we could not analyze their mutation in most of the WGS (**Supplementary Fig. 7**).

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, such as DNA structures, and sequence motif, influence the mutation rate (8, 9, 26, 27). However, little is known about factors that influence the mutability of microsatellites. We first analyzed the replication timings and microsatellite mutation rates (proportion of mutated samples for a microsatellite) (**Fig. 2a-d**). The late-replicating regions had lower mutation rates for insertions and deletions in MSS samples and insertions in MSI samples (**Fig. 2abc**). However, an inverse 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 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 (26), while G-quadruplexes can cause replication

errors (27). The impact of DNA shapes is not well known, but one DNA shape parameter (ORChID2) has 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. 8, 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 also 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**), indicating that genetic variations and somatic mutations are influenced by the same factors.

Highly-mutated microsatellites

We compared mutability of each microsatellite between the MSI and MSS samples, and selected the top 20 highly-mutated microsatellites with the highest mutation rates (proportion of mutated samples for a microsatellite) (**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**). Although the efficiency of these markers should be evaluated by an independent and larger cohort, we consider that they have a technical advantage over known MSI markers in availability in whole genome sequencing and combinations of the markers can be used as a new marker set.

Genes with large number of mutated microsatellites

To find genes with high mutation rates (proportion of mutated samples), 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 compared to 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 largest number of mutated microsatellite (Fisher's exact test; *q-value* = 6.5×10^{-15}, odds ratio = 65.1) in liver cancer (LI) (**Supplementary Fig. 9**). A previous study suggested that some cell linage-specific highly-expressed genes, including *ALB* in liver, had recurrent short indels (28). This result is consistent with the previous 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

To compare recurrently mutated microsatellites in the coding regions between MSS and MSI samples, we calculated the proportion of mutated samples for each microsatellite in coding regions (**Fig 4, Supplementary Table 7**). Microsatellites or repeat sequences 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 (29). In the MSS samples, microsatellites or repeat sequences in *APC* and *TCF12* were 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, 30), our analysis identified new genes with recurrently mutated microsatellites.

Of these, the *GINS1* gene encodes a subunit of DNA replication complex (31). *MBD4* has been reported to contribute to tumorigenesis and work as a modifier of MMR-deficient cancer (32). *BLM* is included in the Cosmic Cancer Gene database and has functions in DNA replication and DNA double-strand break repair (33). These results suggest that mutations in microsatellites can work as driver events.

Mutation in mismatch repair (MMR) and proof-reading 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 could 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 that *PMS2* could have a lower impact on DNA mismatch repair deficiency or MSI (34).

Most MSI samples had larger numbers of somatic SNVs (**Fig. 1c**), due to 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) (35) (**Supplementary Fig. 10, 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. 11, Supplementary Table 9**). The PACWG

signature analysis detected 49 SBS, 11 DBS, and 17 ID signatures (36). 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). Except for DBS3 and DBS8, most of these mutational signatures have been reported to be associated with tumors having defective DNA mismatch repair (36). We found DBS3 and DBS8 to be 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 (36). 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 cancers are known to show specific immune reactions such as Crohn-like reaction and diffused infiltration of lymphocytes in pathology (37), 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 to 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 MSS tumors (median, 11; *P-value* = 5.9×10^{-20}) (**Supplementary Fig. 12**), 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 indels of the microsatellites.

Discussion

Due to the clinical importance of MSI phenotypes, exome sequencing and small-scale WGS studies were performed for the MSI samples (13, 14, 20, 30). 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, detailed analysis for factors that influence mutation rate and validation of the selected microsatellite marker sets in independent cohorts was limited. Here, we performed an analysis of microsatellite mutations in the largest whole genome sequencing cohort with the largest number of microsatellites to date, so as to characterize microsatellite mutations and MSI tumors at the genome-wide level.

To identify microsatellite regions in the human genome, we used results from three software (Tandem repeat finder, MS detector, and MISA) (16-18). After filtering, we obtained 9,292,677 microsatellites in the genome for analysis. Among the methods for detecting microsatellites, the 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 (38). However, current short-reads sequencing technologies are unable to analyze longer microsatellites. Indeed, we could not obtain sufficient number of reads for Bethesda markers (**Supplementary Fig. 7**). We therefore decided to prioritize shorter repeats as the main targets for the current WGS study. Alternatively, in this study, we detected highly-mutated short microsatellites (**Supplementary Table 5**) and they could be useful for clinical diagnosis of MSI with current short read technologies.

The analysis of WGS and the validation study in an independent cohort found 20 novel microsatellite markers (**Fig. 3**), which can be used to predict tumors with high mutational burden. 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. 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, instead of indels and microsatellites (**Supplementary Fig. 10**). Therefore, analysis of mutations in the *POLE* gene in MSS samples can identify more tumors with high mutational burdens (39).

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 previous studies (20). 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 correlated with the insertions and deletions of MSS samples, as well as 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, and higher Slide had higher insertion rates (**Fig. 2e-h, Supplementary Fig. 8**). Microsatellites with higher HelT, higher Opening, higher ProT, higher Roll, and lower Slide had higher deletion rates (**Fig. 2e-h, Supplementary Fig. 8**). 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. 10, 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 microsatellites, 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 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://rohsdb.cmb.usc.edu/GBshape/

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

436 Cosmic cancer genes; https://cancer.sanger.ac.uk/cosmic

437 R; https://www.r-project.org

438

439 **Author Contributions**

440 Study design: A.F., S.G.R., M.R.S., and H.N. Data analysis: A.F., M.F., H.T., Y.S., S.M.

441 S.I. and H.N. Molecular analysis: K.M., A.O-S., K.N. and H.N. S. Manuscript writing:

442 A.F., M.F. and H.N.

443

444

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 microsatellites, we first defined the repeat unit of each microsatellite. We divided each region by 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 (40) 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. The type of microsatellites was defined by length of microsatellite region in the reference genome and repeat unit. Microsatellites were categorized by length (6-9, 10-19, 20-29, 30-39, and 40-100bp) and repeat unit and error rates were estimated for each category (see **Supplementary Table 2**). 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 repeats 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 microsatellites that were covered by ≥ 15 reads in both the cancer and matched normal samples. Additionally, we selected somatic indels with variant allele frequencies in cancer ≥ 0.15 and number of support reads in cancer ≥ 2 and ≤ 1 in normal samples.

Estimation of false discovery rate

We randomly selected 29 somatic MS mutations detected in liver cancer samples RK001, RK249 and RK308 and performed validation with a previously reported method (41). 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 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.

Threshold determination for MSI

Since the mutation status of conventional MSI markers could not be obtained (see **Supplementary Fig. 7**), we needed to determine the threshold value to select MSI. We first excluded colorectal, stomach and uterus cancers (n=186) from all samples (n=2717). We assumed that the other cancers contained negligible number of MSI samples. We then calculated the average and standard deviation of the mutation rates. We also assumed that the distribution of the mutation rates follows a normal distribution with the obtained

average and standard deviation. The 99.99th percentile of the normal distribution was 0.0254. Therefore, we adapted 0.03 (slightly conservative value from the 99.99th percentile) for the criteria for MSI in genome-wide level, and 31 samples were defined as MSI in this study. No colorectal, stomach or uterus cancers were used to determine the threshold value, however, the value still gave a reasonable grouping for colorectal, stomach and uterus cancers (please see the **Fig. 1 (b)**).

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 (36). 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 (42).

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 microsatellites and total number of mutated microsatellites in introns and exons in each gene. We also counted the total number of analyzed microsatellites and total number of mutated microsatellites in the entire genome in MSS samples for each cancer type; (total number of mutated microsatellites in gene *i* in all MSS samples in cancer *j*)/(total number of un-mutated 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 un-mutated microsatellite in entire genome *i* in all MSS samples in cancer *j*) and these were compared with Fisher's exact test. Multiple testing adjustment was done using Benjamini and Hochberg's FDR method (42), and from this analysis, we could obtain genes with larger numbers of mutated microsatellites compared to the entire genome.

Analysis of epigenetic factors on mutability of microsatellite

To find the factors that influence the mutability of microsatellites, we considered replication timing, nuclear lamina binding region, G-quadruplexes, and predicted DNA shapes (See URLs section). For the replication timing, we downloaded data of HepG2, K562, MCF-7, SK-N-SH, and GM12878 cells. We averaged the replication timing within 1Mbp bins for each cell line and bins with standard deviation ≤ 15 were used for the

577 analysis. Presence or absence of nuclear lamina binding region and G-quadruplexes
578 within ±1000bp from the start and end of microsatellites were examined. The predicted
579 DNA shapes (Buckle, HelT, minor groove width (MGW), ORChID2, Opening, ProT,
580 Rise, Roll, Shear, Shift, Slide, Stagger, Stretch and Tilt) of ±5bp from the start and end
581 of each microsatellite were used for the analysis. We performed a multiple regression
582 analysis of the parameters with lm function of R software, and parameter selection was
583 done with step() function.
584
585 *Prediction of neo-antigens*
586 HLA genotyping from WGS data were generated as part of the PCAWG project. Somatic
587 point mutations, non-MS indels detected by the PCAWG project, and MS indels detected
588 by our method were combined and annotated using ANNOVAR. Mutant peptides of
589 length 8–11 residues were assessed for their binding affinity (IC_{50}) to the HLA class I of
590 matched patients using NetMHCpan-3.0 (43). Mutant peptides of IC_{50} < 50 nM were
591 predicted as neo-antigens.
592

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* $= 1.6\times10^{-23}$. 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* $< 1.0\times10^{-200}$. MSI samples are shown in red.

Figure 2. Analysis of mutation rate of each microsatellite. The mutation rate of 198,578 informative microsatellites was analyzed. **(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* $< 1\times10^{-200}$), **(b)** Insertions in MSI samples (χ^2 test; *p-value* $< 1\times10^{-200}$), **(c)** Deletions in MSS samples (χ^2 test; *p-value* $< 1\times10^{-200}$), **(d)** Deletions in MSI samples (χ^2 test; *p-value* $< 1\times10^{-200}$). **(e-h)** Association of DNA shape with mutation rate. Top 1000 A/T microsatellites with 10-30bp 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. In this figure, we divided the microsatellites with mutation rates (0-0.001, 0.001-0.003, and > 0.003 for MSS, and 0-0.4, 0.4-0.8, and > 0.8 for MSI), and showed the DNA shape values. The arrows show base positions with significant association between the DNA shape values and mutation rates (**Supplementary Table 5**). **(i)** Germline variation and somatic mutation rates of

626 microsatellites. The amount of germline variation was estimated using heterozygosity of
627 normal tissues. X-axis; the proportion of mutated samples in each microsatellite (somatic
628 mutation rate). Y-axis; heterozygosity of each microsatellite. Both rates were positively
629 correlated (Pearson's product-moment correlation; $r = 0.31$, $p\text{-}value < 1.0\times10^{-100}$).
630
631 **Figure 3.** Highly-mutated microsatellite markers. Result of clustering analysis with the
632 top 20 microsatellites. **(a)** CR (Colon/Rectum) cancer, **(b)** ST (Stomach) cancer, **(c)** UT
633 (Uterus) cancer. **(d)** Result of validation study in the independent colon cancer cohort
634 (n=36). MSI was defined by BAT25 or BAT26 (upper panel). Twelve new microsatellite
635 markers (middle panel) were successfully amplified and analyzed for indel mutations by
636 MiSeq (**Supplementary Fig. 12**). Fourteen highly-mutated microsatellites or
637 homopolymers in coding regions were analyzed (lower panel). These mutations occurred
638 specifically in MSI colorectal cancers. Mutation status are shown in brown (mutated) and
639 white (WT; wild type) in the panel. Number of mutated samples in the MSI samples are
640 shown in bar plot (right).
641
642 **Figure 4.** Proportion of mutated samples in coding microsatellites. x-axis; proportion of
643 mutated samples in the MSI samples. y-axis; proportion of mutated samples in the MSS
644 samples. Cosmic cancer genes are shown in red.
645
646 **Figure 5.** Mutation in mismatch repair (MMR) and proof-reading genes. Mutated genes,
647 and mutation rate of microsatellites are shown.
648
649 **Figure 6**. Comparison of mutational signatures between the MSI and MSS samples. **(a)**
650 CR (Colon/Rectum) cancer, **(b)** ST (Stomach) cancer, **(c)** UT (Uterus) cancer. Signatures
651 showing significant difference between the MSI and MSS samples are shown in
652 rectangles in the legends (Wilcoxon signed-rank test, $q\text{-}value < 0.05$).
653

654 **References**

655

656 1. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I,
657 Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB,
658 Morganella S, Aure MR, Lingjaerde OC, Langerod A, Ringner M, Ahn SM, Boyault
659 S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens
660 JA, Gerstung M, Hooijer GK, Jang SJ, Jones DR, Kim HY, King TA, Krishnamurthy
661 S, Lee HJ, Lee JY, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporte
662 I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodriguez-Gonzalez FG, Romieu
663 G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J,
664 Tommasi S, Treilleux I, Van den Eynden GG, Vermeulen P, Vincent-Salomon A,
665 Yates L, Caldas C, van't Veer L, Tutt A, Knappskog S, Tan BK, Jonkers J, Borg A,
666 Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S,
667 Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver
668 MJ, Martens JW, Borresen-Dale AL, Richardson AL, Kong G, Thomas G, Stratton
669 MR (2016) Landscape of somatic mutations in 560 breast cancer whole-genome
670 sequences. Nature 534: 47-54
671 2. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, Patch
672 AM, Kakavand H, Alexandrov LB, Burke H, Jakrot V, Kazakoff S, Holmes O,
673 Leonard C, Sabarinathan R, Mularoni L, Wood S, Xu Q, Waddell N, Tembe V,
674 Pupo GM, De Paoli-Iseppi R, Vilain RE, Shang P, Lau LMS, Dagg RA, Schramm
675 SJ, Pritchard A, Dutton-Regester K, Newell F, Fitzgerald A, Shang CA, Grimmond
676 SM, Pickett HA, Yang JY, Stretch JR, Behren A, Kefford RF, Hersey P, Long GV,
677 Cebon J, Shackleton M, Spillane AJ, Saw RPM, Lopez-Bigas N, Pearson JV,
678 Thompson JF, Scolyer RA, Mann GJ (2017) Whole-genome landscapes of major
679 melanoma subtypes. Nature 545: 175-180
680 3. Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J,
681 Ehrenberger T, Grobner S, Segura-Wang M, Zichner T, Rudneva VA, Warnatz HJ,
682 Sidiropoulos N, Phillips AH, Schumacher S, Kleinheinz K, Waszak SM, Erkek S,
683 Jones DTW, Worst BC, Kool M, Zapatka M, Jager N, Chavez L, Hutter B, Bieg M,
684 Paramasivam N, Heinold M, Gu Z, Ishaque N, Jager-Schmidt C, Imbusch CD,
685 Jugold A, Hubschmann D, Risch T, Amstislavskiy V, Gonzalez FGR, Weber UD,
686 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

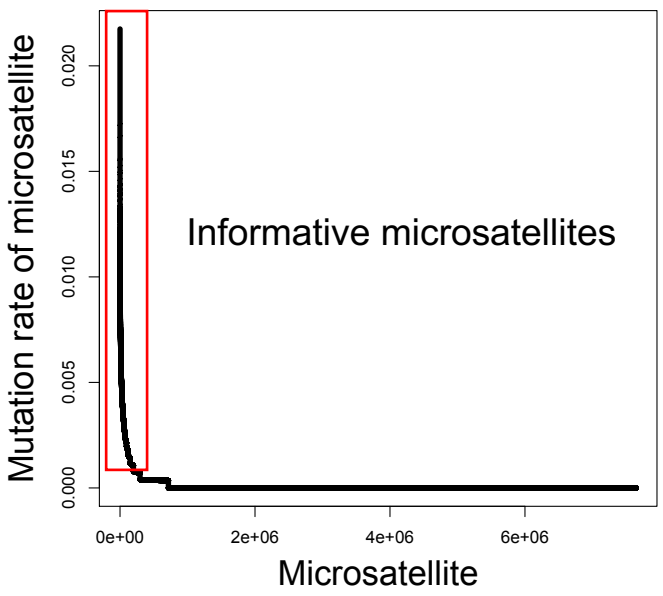
753 18. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences.
754 Nucleic Acids Res 27: 573-80
755 19. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki
756 M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T (2010)
757 Whole-genome sequencing and comprehensive variant analysis of a Japanese
758 individual using massively parallel sequencing. Nat Genet 42: 931-6
759 20. Maruvka YE, Mouw KW, Karlic R, Parasuraman P, Kamburov A, Polak P,
760 Haradhvala NJ, Hess JM, Rheinbay E, Brody Y, Koren A, Braunstein LZ, D'Andrea
761 A, Lawrence MS, Bass A, Bernards A, Michor F, Getz G (2017) Analysis of
762 somatic microsatellite indels identifies driver events in human tumors. Nat
763 Biotechnol 35: 951-959
764 21. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD (2018) Pan-cancer analysis of
765 whole genomes. bioRxiv
766 22. Goldman et al. (2018) Online resources for PCAWG data exploration, visualization,
767 and discovery. bioRxiv
768 23. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, Nakagawa H,
769 Sotamaa K, Prior TW, Westman J, Panescu J, Fix D, Lockman J, Comeras I, de la
770 Chapelle A (2005) Screening for the Lynch syndrome (hereditary nonpolyposis
771 colorectal cancer). N Engl J Med 352: 1851-60
772 24. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, Comeras I, La
773 Jeunesse J, Nakagawa H, Westman JA, Prior TW, Clendenning M, Penzone P,
774 Lombardi J, Dunn P, Cohn DE, Copeland L, Eaton L, Fowler J, Lewandowski G,
775 Vaccarello L, Bell J, Reid G, de la Chapelle A (2006) Screening for Lynch
776 syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer
777 patients. Cancer Res 66: 7810-7
778 25. Arai T, Sakurai U, Sawabe M, Honma N, Aida J, Ushio Y, Kanazawa N, Kuroiwa
779 K, Takubo K (2013) Frequent microsatellite instability in papillary and solid-type,
780 poorly differentiated adenocarcinomas of the stomach. Gastric Cancer 16: 505-12
781 26. Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD (2012) A
782 genome-wide analysis of common fragile sites: what features determine
783 chromosomal instability in the human genome? Genome Res 22: 993-1005
784 27. Lemmens B, van Schendel R, Tijsterman M (2015) Mutagenic consequences of a
785 single G-quadruplex demonstrate mitotic inheritance of DNA replication fork

barriers. Nat Commun 6: 8909

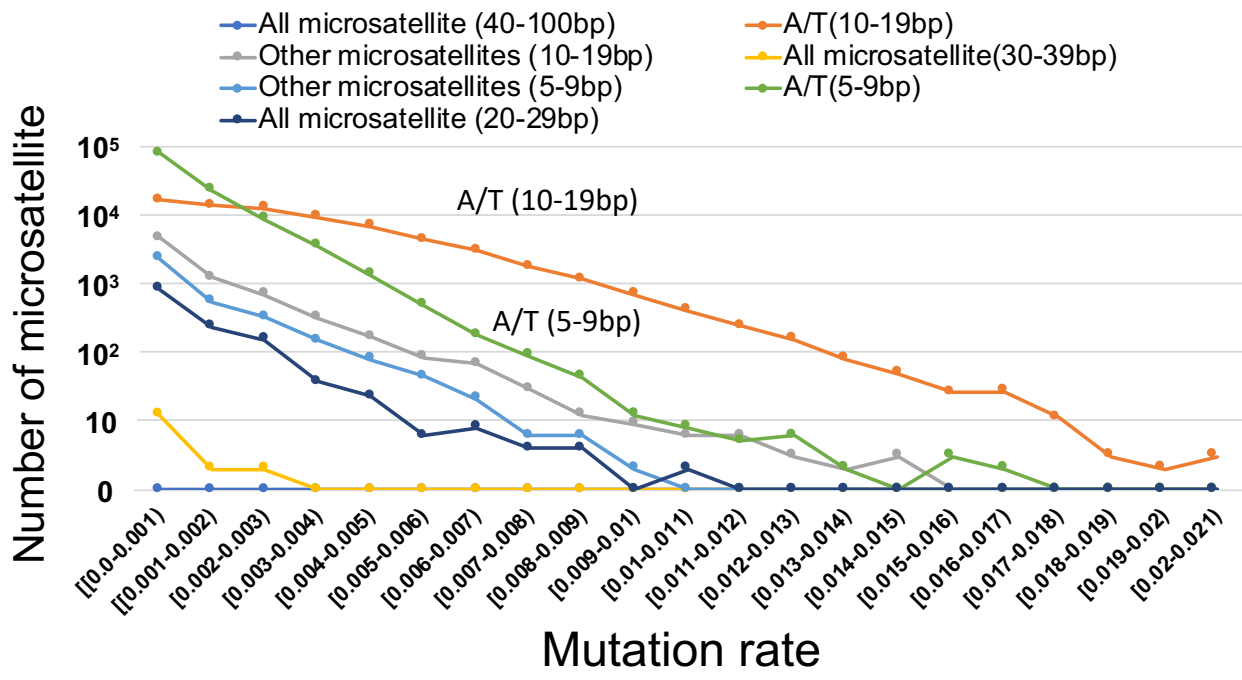
28. Imielinski M, Guo G, Meyerson M (2017) Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. Cell 168: 460-472 e14
29. Albacker LA, Wu J, Smith P, Warmuth M, Stephens PJ, Zhu P, Yu L, Chmielecki J (2017) Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion. PLoS One 12: e0176181
30. 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
31. 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
32. 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
33. 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
34. 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
35. 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

819 36. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington
820 KR, Gordenin DA, Bergstrom E, Lopez-Bigas N, Klimczak LJ, McPherson JR,
821 Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG,
822 Stratton MR (2018) The Repertoire of Mutational Signatures in Human Cancer.
823 bioRxiv
824 37. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, Fishel
825 R, Lindor NM, Burgart LJ, Hamelin R, Hamilton SR, Hiatt RA, Jass J, Lindblom A,
826 Lynch HT, Peltomaki P, Ramsey SD, Rodriguez-Bigas MA, Vasen HF, Hawk ET,
827 Barrett JC, Freedman AN, Srivastava S (2004) Revised Bethesda Guidelines for
828 hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite
829 instability. J Natl Cancer Inst 96: 261-8
830 38. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S,
831 Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of
832 human mutation based on microsatellites. Nat Genet 44: 1161-5
833 39. Mehnert JM, Panda A, Zhong H, Hirshfield K, Damare S, Lane K, Sokol L, Stein
834 MN, Rodriguez-Rodriquez L, Kaufman HL, Ali S, Ross JS, Pavlick DC, Bhanot G,
835 White EP, DiPaola RS, Lovell A, Cheng J, Ganesan S (2016) Immune activation
836 and response to pembrolizumab in POLE-mutant endometrial cancer. J Clin Invest
837 126: 2334-40
838 40. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-
839 64.
840 41. Schuelke M (2000) An economic method for the fluorescent labeling of PCR
841 fragments. Nat Biotechnol 18: 233-4
842 42. Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and
843 powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300 (1995).
844 43. Nielsen M, Andreatta M (2016) NetMHCpan-3.0; improved prediction of binding to
845 MHC class I molecules integrating information from multiple receptor and peptide
846 length datasets. Genome Med 8: 33
847
848
849

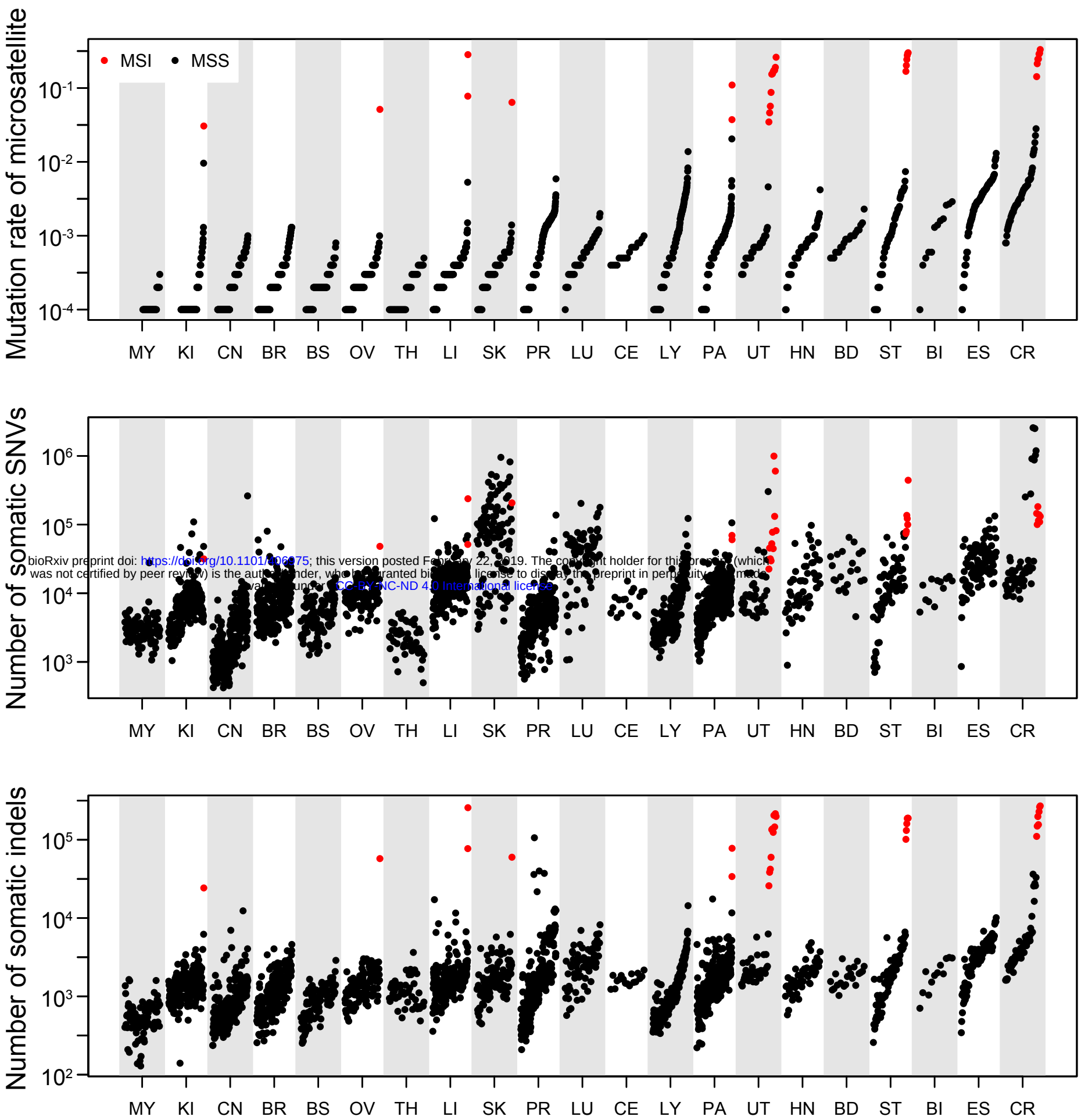
(a)



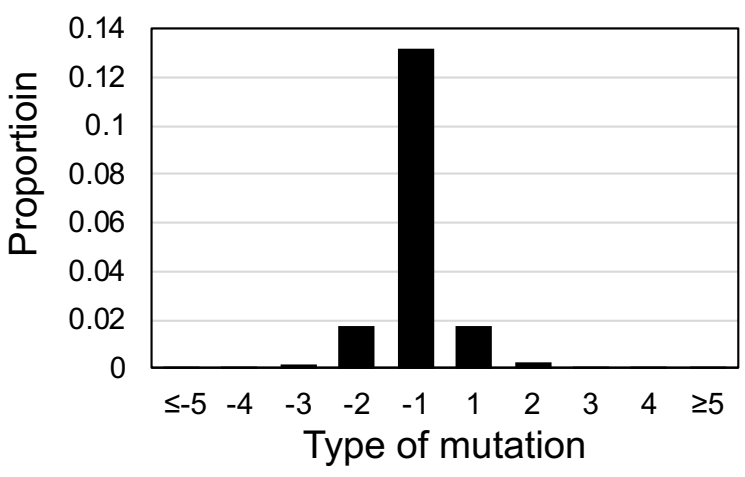
(b)



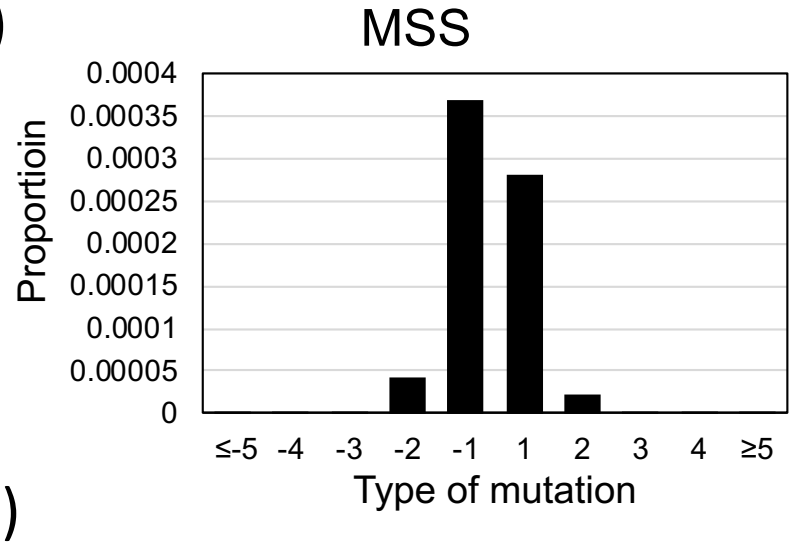
(c)



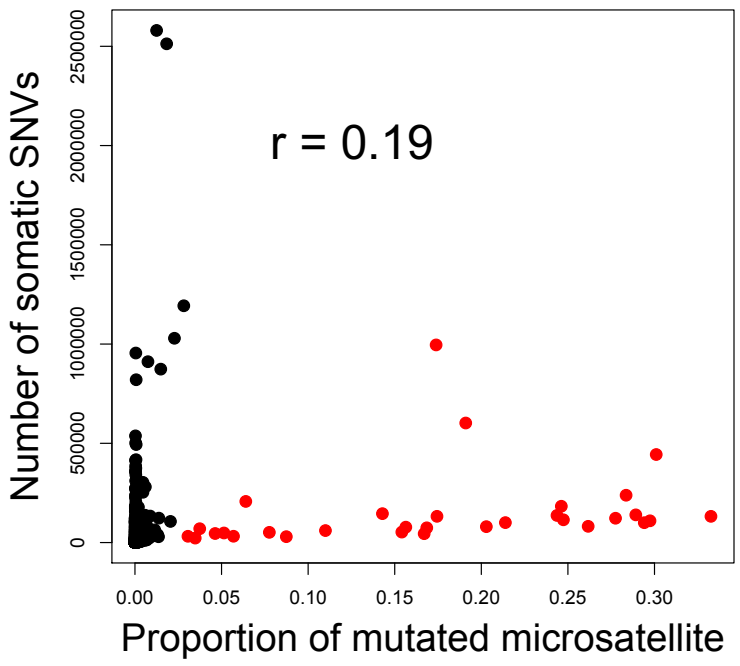
(d)



(e)



(f)



(g)

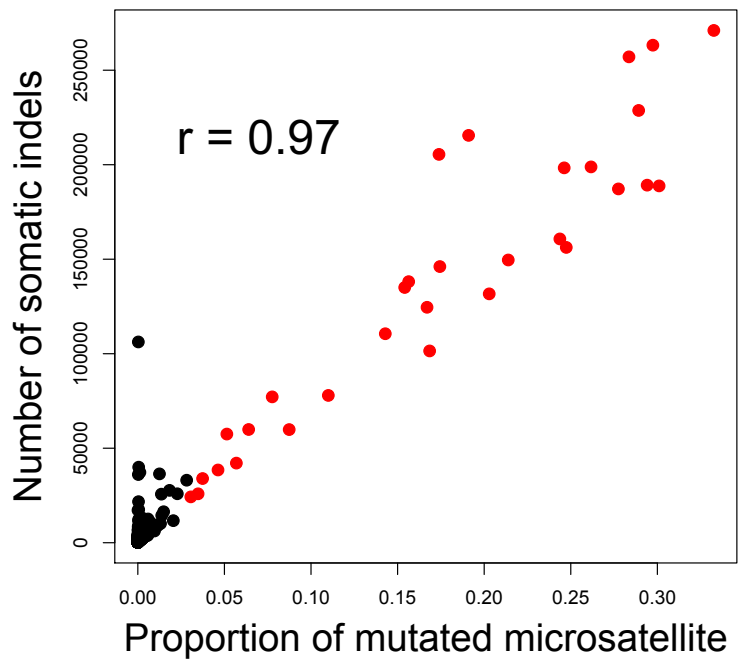


Fig. 1

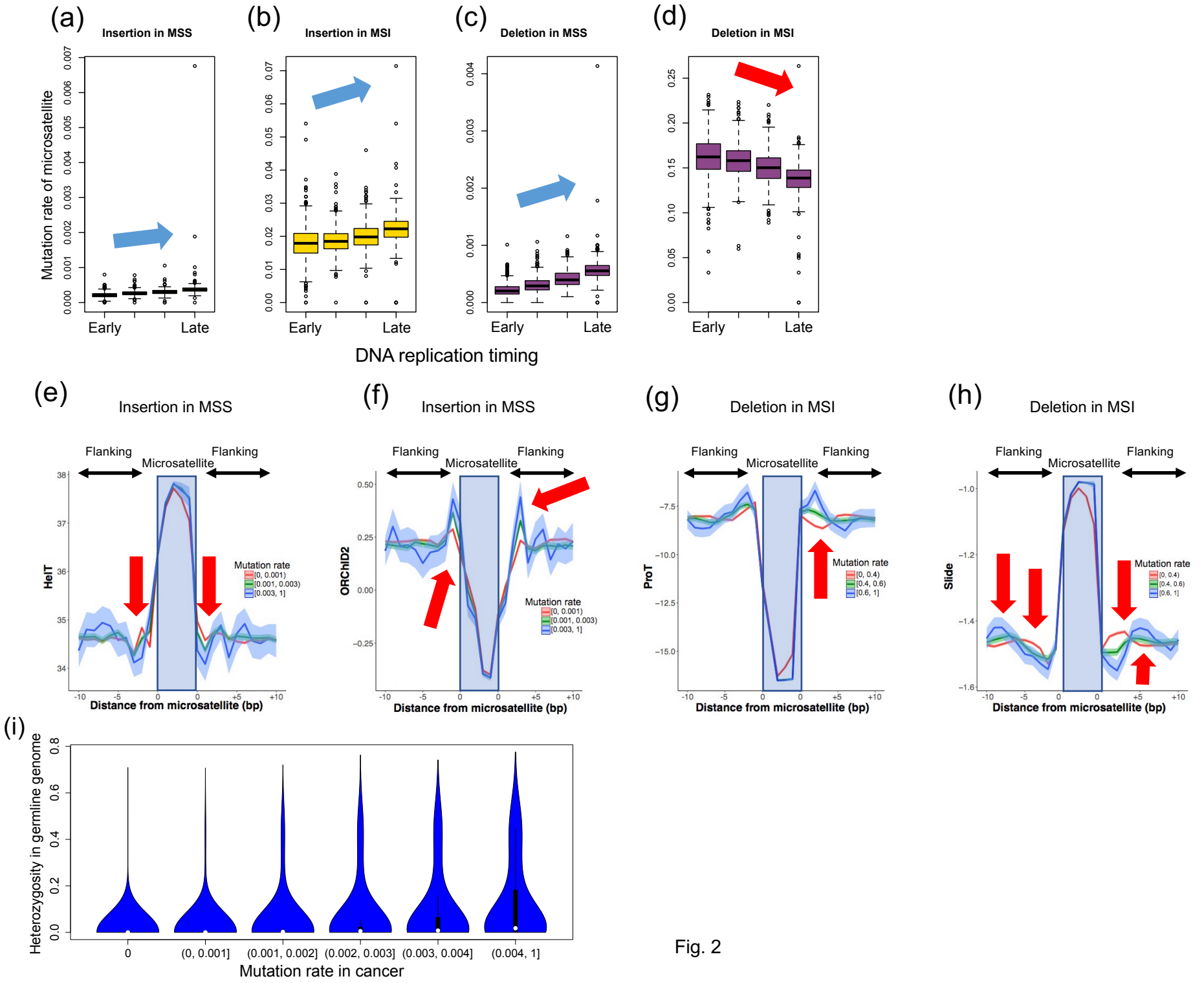

(a)
Insertion in MSS
(b)
Insertion in MSI
(c)
Deletion in MSS
(d)
Deletion in MSI
Mutation rate of microsatellite
Early
Late
DNA replication timing
(e)
Insertion in MSS
(f)
Insertion in MSS
(g)
Deletion in MSI
(h)
Deletion in MSI
Flanking
Microsatellite
Flanking
HelT
ORChID2
ProT
Slide
Mutation rate
[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

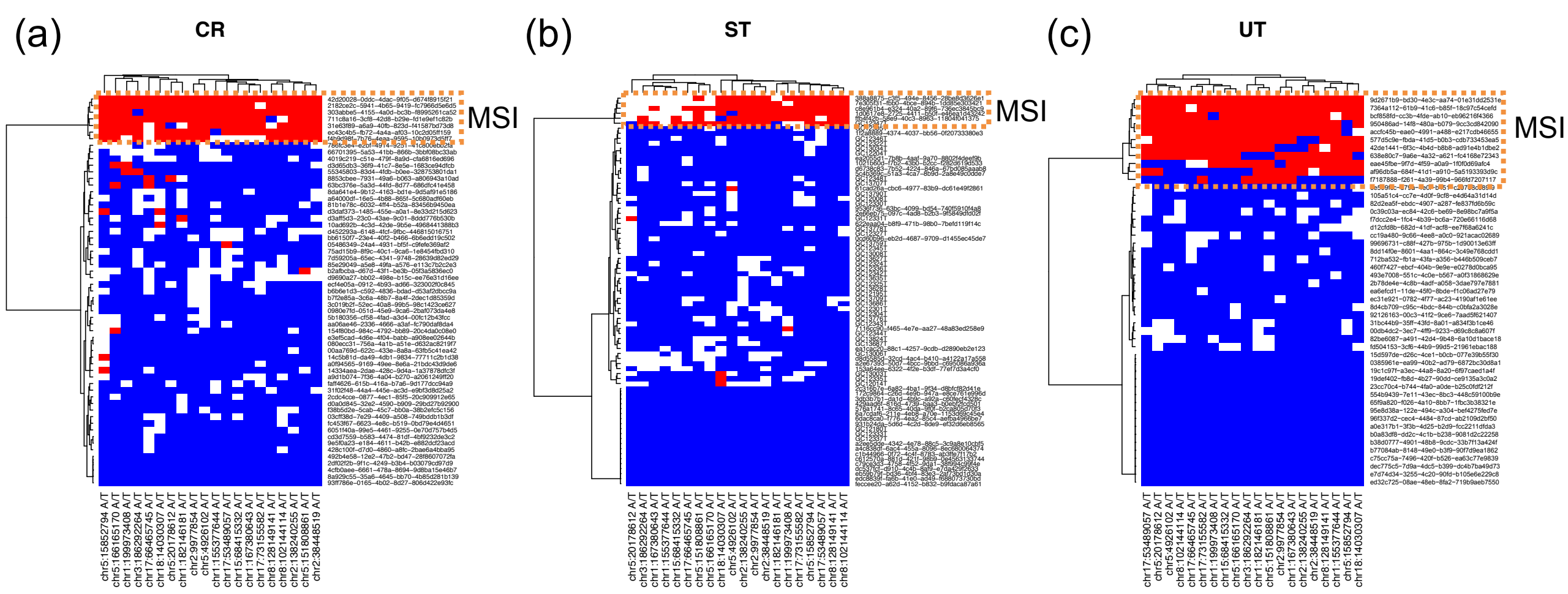
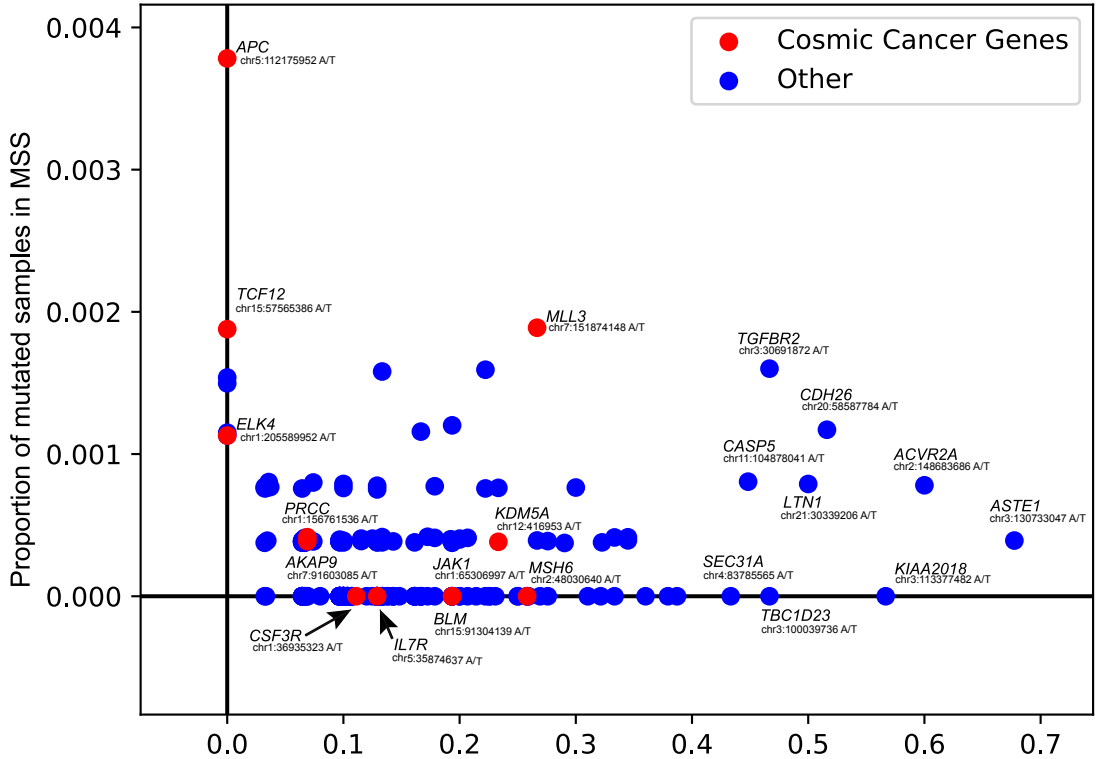


Fig. 3

MSS specific



MSI specific

Fig. 4

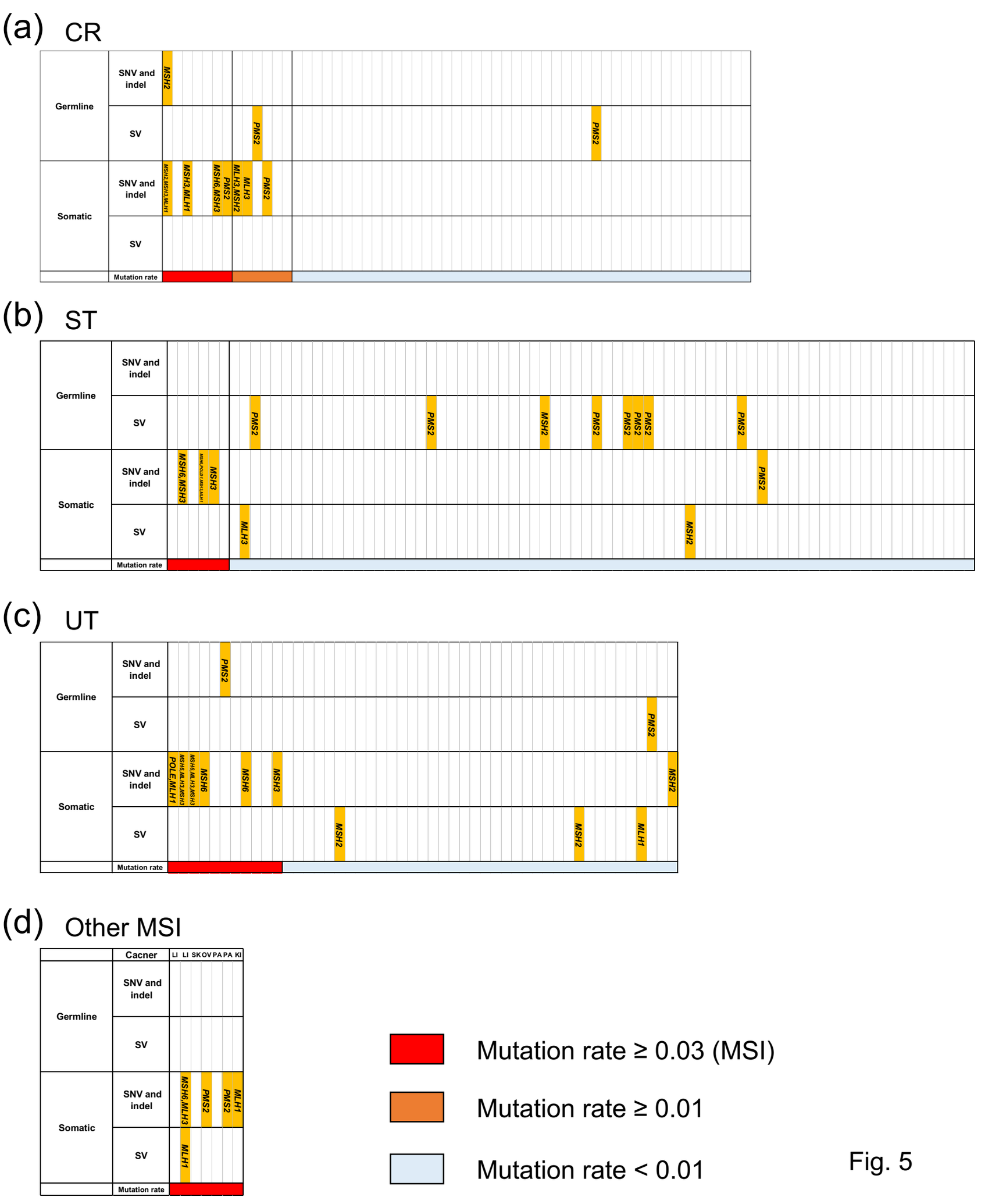

(a) CR
Germline
SNV and indel
SV
Somatic
SNV and indel
SV
Mutation rate
MSH2
PMS2
PMS2
MSH2,MSH3,MLH1
MSH3,MLH1
MSH6,PMS2
MSH6,MSH3
MLH3,MSH2
MLH3
PMS2
(b) ST
Germline
SNV and indel
SV
Somatic
SNV and indel
SV
Mutation rate
PMS2
PMS2
MSH2
PMS2
PMS2
PMS2
PMS2
PMS2
MSH6,MSH3
MSH3
MSH6,POLD1,MSH3,MLH1
PMS2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

Fig. 5

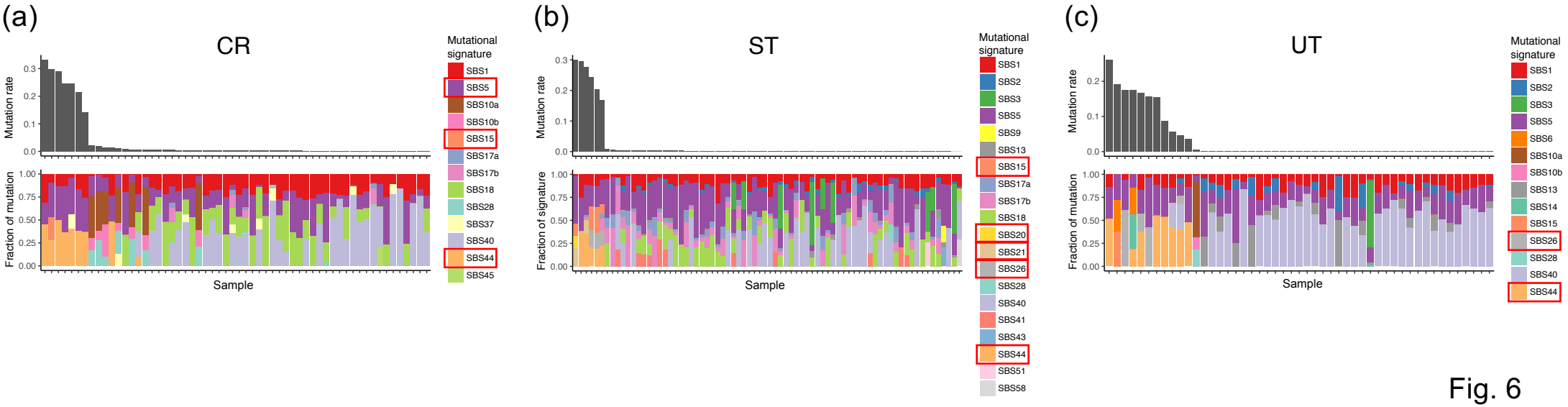

(a)
CR
Mutation rate
Fraction of mutation
Sample
Mutational signature
SBS1
SBS5
SBS10a
SBS10b
SBS15
SBS17a
SBS17b
SBS18
SBS28
SBS37
SBS40
SBS44
SBS45
(b)
ST
Mutation rate
Fraction of signature
Sample
Mutational signature
SBS1
SBS2
SBS3
SBS5
SBS9
SBS13
SBS15
SBS17a
SBS17b
SBS18
SBS20
SBS21
SBS26
SBS28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