Cell

The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes

Tae-Min Kim,[1,2] Peter W. Laird,[3] and Peter J. Park[1,4,*]
[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[2]Cancer Evolution Research Center, College of Medicine, The Catholic University of Korea, Seoul 137-701, Korea
[3]USC Epigenome Center, Department of Surgery and of Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[4]Informatics Program, Children's Hospital Boston and Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA
*Correspondence: peter_park@harvard.edu
http://dx.doi.org/10.1016/j.cell.2013.10.015

SUMMARY

Microsatellites—simple tandem repeats present at millions of sites in the human genome—can shorten or lengthen due to a defect in DNA mismatch repair. We present here a comprehensive genome-wide analysis of the prevalence, mutational spectrum, and functional consequences of microsatellite instability (MSI) in cancer genomes. We analyzed MSI in 277 colorectal and endometrial cancer genomes (including 57 microsatellite-unstable ones) using exome and whole-genome sequencing data. Recurrent MSI events in coding sequences showed tumor type specificity, elevated frameshift-to-inframe ratios, and lower transcript levels than wild-type alleles. Moreover, genome-wide analysis revealed differences in the distribution of MSI versus point mutations, including overrepresentation of MSI in euchromatic and intronic regions compared to heterochromatic and intergenic regions, respectively, and depletion of MSI at nucleosome-occupied sequences. Our results provide a panoramic view of MSI in cancer genomes, highlighting their tumor type specificity, impact on gene expression, and the role of chromatin organization.

INTRODUCTION

About 15% of sporadic colorectal cancers (CRC) harbor widespread alterations in the length of microsatellite (MS) sequences, known as microsatellite instability (MSI) (Boland and Goel, 2010). Sporadic MSI CRC tumors display unique clinicopathological features, including near-diploid karyotype, higher frequency in older populations and in females, and a better prognosis (de la Chapelle and Hampel, 2010; Popat et al., 2005). MSI is known to occur due to a defective DNA mismatch repair (MMR) system with key MMR genes inactivated by various mechanisms such as germline mutation in *MSH2* or *MLH1* in most Lynch syndrome cases (Bronner et al., 1994; Leach et al., 1993) and epigenetic silencing of *MLH1* in most sporadic cases (Herman et al., 1998; Veigl et al., 1998). The DNA slippage within coding sequences can induce frameshifting mutations that result in the production of truncated, functionally inactive proteins. For CRC genomes, cancer-related genes frequently targeted by MSI (e.g., *TGFBR2*, *ACVR2A*, and *BAX*) have been studied (Jung et al., 2004; Markowitz et al., 1995; Rampino et al., 1997). MSI is also present in other tumors, such as in endometrial cancer (EC) of the uterus, the most common gynecological malignancy (Duggan et al., 1994). The same reference Bethesda panel originally developed to screen an inherited genetic disorder (Lynch syndrome) (Umar et al., 2004) is currently applied to test MSI for CRCs and ECs. However, the genes frequently targeted by MSI in CRC genomes rarely harbor DNA slippage events in EC genomes (Gurin et al., 1999), and it is largely unknown whether MS-unstable EC genomes have similar molecular origins or functional consequences as CRC genomes.

In this study, we utilize the exome and whole-genome sequencing data for CRC and EC genomes from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2013) to profile the genomic landscape of MSI in these two tumor types, including the patterns of single-nucleotide variations (SNV) in MMR pathways, a comprehensive catalog of genomic loci with frequent MSI, the genomic distribution and sequence properties of the affected microsatellites, and correlations with other genomic and epigenetic features.

RESULTS

The Mutational Spectrum of Exome-wide MSI in Cancer Genomes

To examine the impact of MSI on protein-coding sequences, we analyzed the exome sequencing data for 147 CRC and 130 EC patients (Table S1 available online). The initial cohorts included 27 CRC/30 EC MSI-H (MSI-high), 23/11 MSI-L (MSI-low), and 97/89 MSS (MS-stable) genomes, as evaluated by the revised Bethesda guidelines (Umar et al., 2004). We used computational methods to identify MS contained within sequencing reads and

to detect significant differences in their lengths between tumor and matched normal genomes (see the Experimental Procedures). Whereas the current Bethesda panel categorization simply classifies cancer genomes into MSI-H, MSI-L, and MSS based on the number of markers altered, our analysis shows that MSI-H genomes show a dramatically higher number of MSI events (median of 290 and 126 MSI events per MSI-H CRC and EC genome, respectively) compared to MSI-L (median of 5 and 2) and MSS (median of 4 and 1) in both cancer types (Figures 1A, 1B, and S1). The difference in the number of MSI events is not significant between MSI-L and MSS ($p = 0.22$ and $p = 0.42$ for CRC and EC; Figure S1). Details on the identified MSI events are in Tables S2 and S3. When corrected for the background distribution of different repeat types in the exome reference set of MS, we observe a depletion of MSI events in coding sequences, likely reflecting purifying selection of mutations involving coding sequences (Figure 1C).

We next examined the relationship between MSI events and SNV mutation rates as well as the mutation status of key MMR genes (Figures 1A and 1B). Our combined mutational profiles highlight three main features. First, we observe the vulnerability of specific MMR genes to different types of somatic mutations as their inactivating mechanism. Although most of the MSI-H CRC and EC genomes harbor transcriptional silencing of *MLH1* by promoter hypermethylation, frameshifting DNA slippage events are the primary inactivating mechanism for *MSH3* and, to a lesser extent, for *MSH6* in MS-unstable CRC and EC genomes. Other MMR genes such as *MSH2*, *PMS1*, and *PMS2* only harbor nonsilent (missense or nonsense) SNVs, mostly in the hypermutated samples. Second, complementary mechanisms of inactivation are observed for some genes. For example, nonsilent SNVs and DNA slippage events are mutually exclusive for both *MSH3* and *MSH6* in MS-unstable genomes, suggesting that these two may be alternative mechanisms for inactivation of those genes (Ciriello et al., 2012). Third, a number of samples show highly elevated SNV mutation rates, most of them harboring missense mutations of *POLE* (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2013), but there is no relationship between SNV mutation rates and MSI. In addition, *POLE*-mutated genomes can be largely classified into two classes depending on the *MLH1* status: MS-unstable genomes (inactivation of *MLH1*) and MS-stable ones (functional *MLH1*). The highly elevated mutation rates are observed for the latter. It is possible that *POLE* mutations in MS-unstable genomes are late events. Alternatively, MSI is sufficient to achieve the phenotypes required by cancer cells in MS-unstable genomes and/or these genomes do not tolerate the additional mutation burden from SNVs. Our observations also highlight the primary role of *MLH1* inactivation in the establishment of an MSI phenotype because *POLE*-mutated genomes with functional *MLH1* maintain the MS stability in the presence of frequent nonsilent SNVs in MMR genes. We observe two *POLE*-mutated MSI-H genomes (1 CRC and 1 EC; arrows in Figure 1) with nonsilent *MLH1* mutations, but not transcriptional silencing of *MLH1*, in which the genomic instability associated with *POLE* mutation might have triggered inactivation of *MLH1*, leading to the MSI phenotype.

Loci Frequently Targeted by MSI Show a Higher Rate of Frameshift Events

For each MSI event, we examined the distribution of changes in the length of the mutant MS allele compared to its germline counterpart. After clustering the MSI events, the heatmap, which mimics the electrophoretic autoradiogram in a conventional MSI study, illustrates the extent of allelic shift for each MSI event (Figures 2A and 2D). Most allelic shifts are deletions, and a higher allelic shift in the length of the mutant allele is more frequent in 3′ UTR than in coding regions. Figure 2A is for MSI events at mononucleotide repeats; a similar pattern is also observed for dinucleotide repeats (Figure S2). We further classified MSI events into low- and high-allelic shift (LAS and HAS, respectively), depending on whether the mode (most frequent value) of the MS allele lengths is equal to its germline length or not. The ratio of LAS/HAS events is higher in coding regions than in 5′/3′ UTRs or noncoding regions (Figures 2B and 2E). An evolutionary model previously proposed (Tsao et al., 2000) suggests that HAS events are more likely to have functional impact than LAS events. Thus, a substantially higher LAS/HAS ratio provides additional evidence for negative selection of MSI events in coding regions.

MSI events on trinucleotide repeats were primarily observed in coding sequences and showed comparable numbers of LAS and HAS events for coding MSI (Figure S2), probably due to their relatively neutral nature (i.e., in-frame) in coding sequences (Metzgar et al., 2000). Thus, we further categorized coding MSI into nontriplet (frameshift) and triplet (in-frame) events, similar to the distinction between nonsynonymous and synonymous SNV mutations (Greenman et al., 2007). The percentage of frameshift and in-frame MSI events is shown with respect to the level of recurrence for CRC and EC genomes (Figures 2C and 2F). MSI events of mononucleotide repeats are largely responsible for this relationship given the predominance of mononucleotide-MSI events (92.4% and 93.0% of total MSI events in CRC and EC genomes). For both tumor types, nonrecurrent coding MSI events show a lower frequency of frameshift MSI events compared to those occurring in noncoding or UTR regions (non-CDS) at a similar recurrence level, concordant with the negative selection of frameshifting MSI events on coding sequences. Importantly, highly recurrent coding MSI events show a higher frameshift-to-inframe ratio compared to nonrecurrent coding MSI (Figures 2C and 2F). This suggests that these nonneutral genomic events may provide selective advantages to the affected clones to overcome the purifying selection on mutations involving coding sequences. Thus, we hypothesize that the genes inactivated by the recurrent frameshift MSI may have tumor-suppressive roles in CRC and EC genomes.

We evaluated the performance of our sequencing-based method by comparing its MSI calls on one of the Bethesda markers (*TGFBR2*, A_{10} homopolymer) with those from the fragment length assay by Sanger sequencing. Sequencing-based MSI screening identified 20 out of 22 *TGFBR2* MSI calls made from Sanger sequencing in 126 CRC genomes without false positives (sensitivity and specificity of 91% and 100%, respectively). For EC genomes, exome and Sanger calls for *TGFBR2* were made only on 3 of 130 genomes, and they were concordant in every case. Examples of one positive and one negative *TGFBR2*

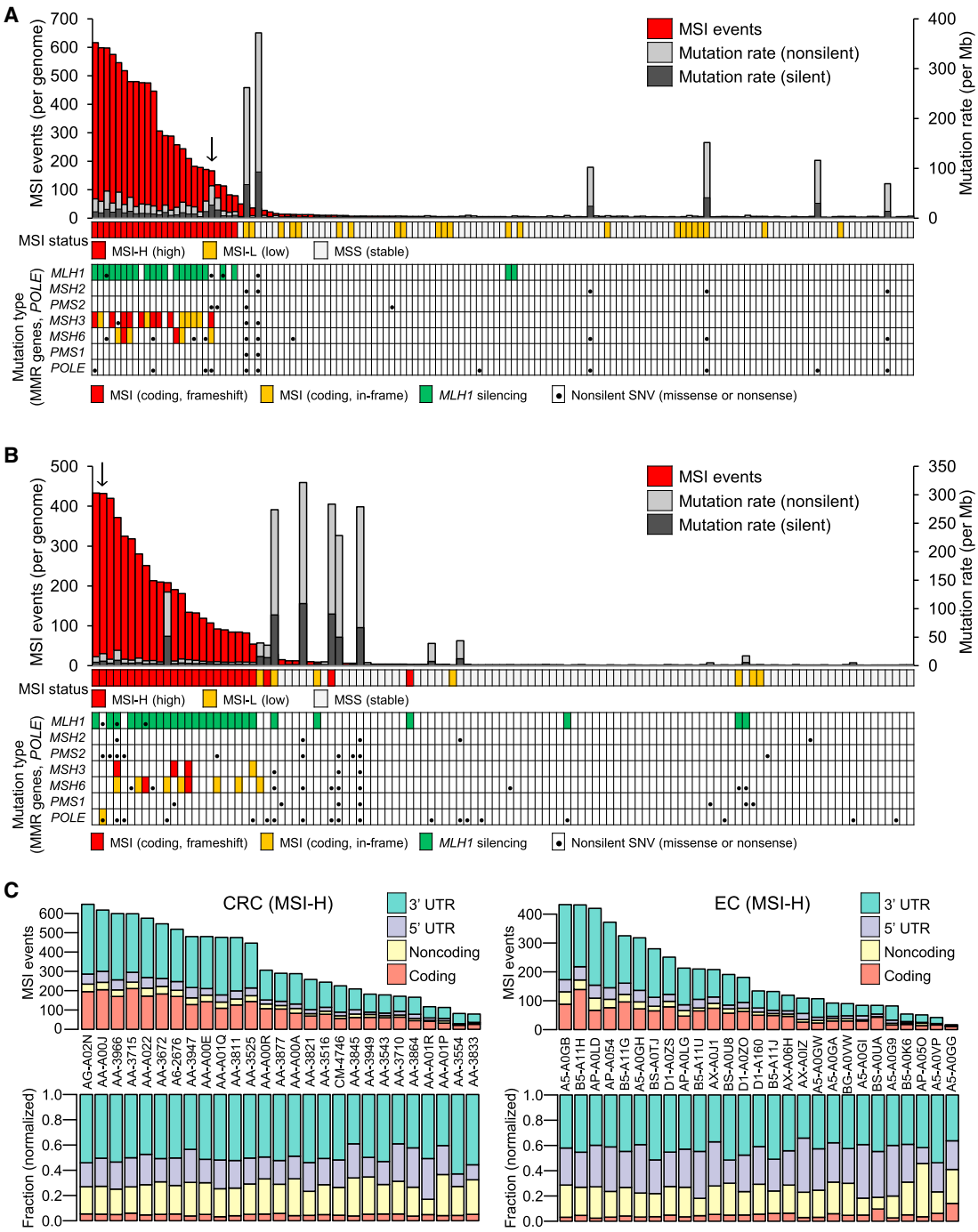

Figure 1. The Mutational Spectrum of MSI Events and MMR Genes in CRC and EC Genomes

(A) The number of MS loci with significant tumor genome-specific DNA slippage events is shown for each of CRC genomes (141 cases with data on *MLH1* promoter hypermethylation are displayed out of 147; see also Figure S1), along with the SNV mutation rate. The samples are sorted in decreasing order of MSI events. The MSI status based on the Bethesda criteria (25 MSI-H, 23 MSI-L, and 93 MSS cases) is noted. The functional status of selected MMR genes and *POLE* are classified into MSI events (frameshift and in-frame), nonsilent point mutations (missense or nonsense), and transcriptional silencing of *MLH1* by hypermethylation. The arrow points to a *POLE*-mutated MSI-H genome with an *MLH1* mutation discussed in the Results.

(B) Similar to (A) for EC genomes (115 cases with *MLH1* promoter hypermethylation data are displayed out of 130).

(C) (Left) For the 27 MSI-H CRC genomes, the numbers of MSI events in the four different categories of genomic regions (coding, noncoding and 5′/3′ UTR) are shown in the upper panel. In the lower panel, the number of MSI events was normalized by the total number of MS in the exome reference set for each category. (Right) Same analysis for MSI-H EC genomes (three samples with <10 MSI were removed).

See also Figure S1 and Tables S1, S2, and S3.

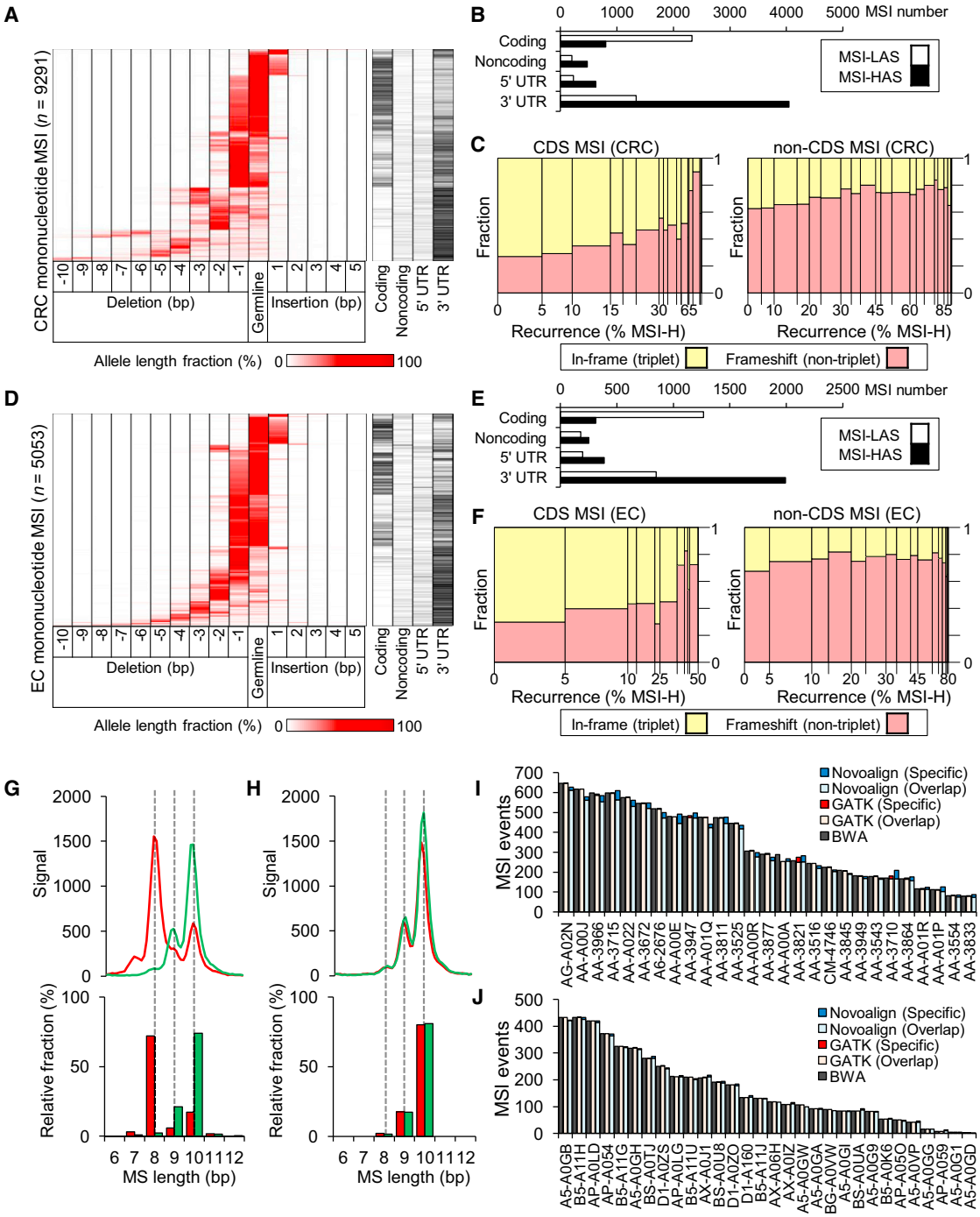

Figure 2. The Distribution of Allelic Shift in MSI Events and the Properties of Recurrent Coding MSI

(A) For MSI events occurring at the mononucleotide MS (y axis; each row) in the CRC genomes, the deviations in the allele lengths (−10 bp to +5 bp) compared to the germline counterparts are shown as normalized allelic fractions in a heatmap (the values in each row add up to 1), clustered by their similarity. The locations of the corresponding MS (coding, noncoding and 5′/3′ UTR) are shown on the right.

(B) MSI events are classified into low- and high-allelic shift (MSI-LAS and MSI-HAS) cases. The graph shows the different frequencies of the two MSI types for the four categories in CRC genomes.

(C) MSI events in the coding sequences (CDS) and non-CDS regions are further classified into frameshift and in-frame mutations for CDS (nontriplet and triplet for non-CDS). The frameshift-to-inframe ratio increases with respect to the level of recurrence (% of MS-unstable genomes harboring the mutation; the width of each bar is proportional to the number of MSI) for CDS MSI events.

(D–F) Similar to the above for EC genomes.

(legend continued on next page)

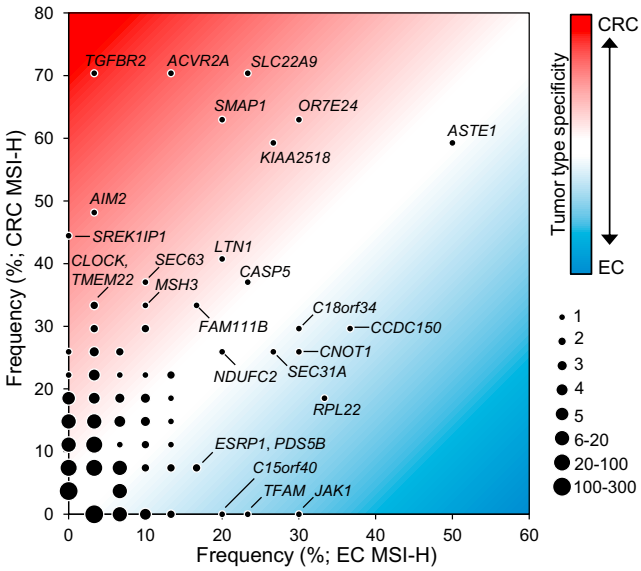

Figure 3. The Genes Harboring Frameshift MSI in CRC and EC Genomes and Tumor Type Specificity

A scatterplot shows the distribution of genes with respect to their frequency of frameshift MSI in CRC and EC genomes. The 27 genes with frameshift MSI in >30% of CRC or in >15% of EC MSI-H genomes are noted. The color gradient indicates the extent of tumor type specificity (red and blue for CRC and EC specificity, respectively). The size of the circles indicates the number of genes with the corresponding MSI frequencies. See also Figure S3 and Table S4.

MSI call are shown in Figures 2G and 2H (see also Figure S2). These results strongly support the robustness of our sequencing-based MSI calls. In the two false-negative cases, the differences observed in the distributions of MS lengths were not statistically significant due to low read coverage. Refinement on the significance threshold may improve sensitivity for the exome-based approach.

Next, we calculated the frequency of frameshift MSI for each gene in the MSI-H tumors (Figure 3). The frequencies in the two tumor types were moderately correlated (R = 0.470), with some loci such as *ASTE1* and *CASP5* showing comparable MSI frequency in both. But we also discovered a substantial number of genes targeted by recurrent frameshift MSI with tumor type specificity. These genes include a few well-known ones such as *ACVR2A* and *TGFBR2* (Markowitz et al., 1995; Parsons et al., 1995; Wang et al., 1995), as well as *MSH3*. Various molecular functions are perturbed by CRC-specific recurrent MSI events, e.g., *SLC22A9* and *TMEM22* encode transport-related molecules, and *SREK1IP1*, *LTN1*, and *SEC63* are related to protein metabolism. Among the novel loci with frequent frameshift MSI such as *SMAP1* and *AIM2*, the potential apoptotic role of *AIM2* has been reported (Fernandes-Alnemri et al., 2009).

Among the novel genes with EC-specific frameshift MSI, MSI events on *JAK1* coding sequences were observed in 30% of MSI-H EC genomes, but none were observed in CRC genomes. Although the protein tyrosine kinase encoded by *JAK1* has been reported as an upstream component of the oncogenic JAK-STAT signaling pathway, whether the locus is frequently subject to MSI or what its functional implication might be was largely unknown. Gene set enrichment analysis (GSEA) revealed that MS-unstable EC genomes harboring the *JAK1* frameshift MSI may have suppression of JAK-STAT signaling, as evidenced by the repressed transcript levels of genes belonging to the pathway and the transcriptional activation of cell-cycle-related genes (Figure S3). *TFAM* also showed EC-exclusive frameshift MSI. Mitochondrial transcription factor A (mtTFA) encoded by *TFAM* has a role in apoptosis and DNA repair (Larsson et al., 1998), and expression of mtTFA was associated with cancer prognosis (Nakayama et al., 2012). EC-specific frameshift MSI events were also observed in *PDS5B*, whose interaction with BRCA2 is required for BRCA2-RAD51-mediated DNA damage repair process (Brough et al., 2012), and in *ESRP1*, whose underexpression is involved in the aberrant splicing pattern during TGFβ-induced epithelial-mesenchymal transformation (Horiguchi et al., 2012). In addition, immune and apoptosis-related gene functions are enriched in genes frequently targeted by frameshift MSI in both tumor types (Table S4).

Bias in Allelic Expression due to MSI Events

To investigate the potential influence of MSI events on the expression level of the affected genes, we compared the differential allelic read counts (wild-type versus mutant alleles) from RNA sequencing (RNA-seq) with those from exome data. A statistically significant bias (false discovery rate [FDR] < 0.05, Fisher's exact test) was observed for 223 and 131 MSI calls in the MS-unstable CRC and EC genomes, corresponding to 16% and 11% of the total MSI calls with a minimum of 10 RNA-seq reads (Table S5). When we categorized these biases into overexpressed MSI ($RNA\text{-}seq_{mutant}/RNA\text{-}seq_{wild\text{-}type} > Exome_{mutant}/Exome_{wild\text{-}type}$) and underexpressed MSI, most of the frameshift MSI were in the underexpressed group in both tumor types (Figures 4A and 4B).

For genes with significant allelic expression biases in multiple samples, the expression changes for transcripts with the mutant alleles were generally in the same direction (33 of 37 genes for CRC and 8 of 14 for EC showed perfect concordance; Figure 4C). For example, the MS alleles with DNA slippage events in the 3′ UTR of *ANTXR1* showed significantly lower transcript levels than the wild-type alleles in all eight CRC genomes. We also compared the transcript levels between the genomes with and without the corresponding MSI (Figure 4D). The expression changes were concordant with the within-sample ratios of over- or underexpression of the mutant allele, with 13 genes showing significant differences.

(G) The distribution of A_{10} homopolymer length on *TGFBR2* locus is shown for one CRC genome with positive MSI calls as measured by Sanger (upper) and exome sequencing data (below).

(H) Similar to AA-2676 as an MSI-negative example.

(I and J) The MSI events per sample are compared to those made after local realignment by GATK or by global realignment by Novoalign for 27 MSI-H CRC (I) and 30 EC genomes (J). Overlap and specific calls are distinguished to those overlapped with BWA-based calls or not, respectively.

See also Figure S2.

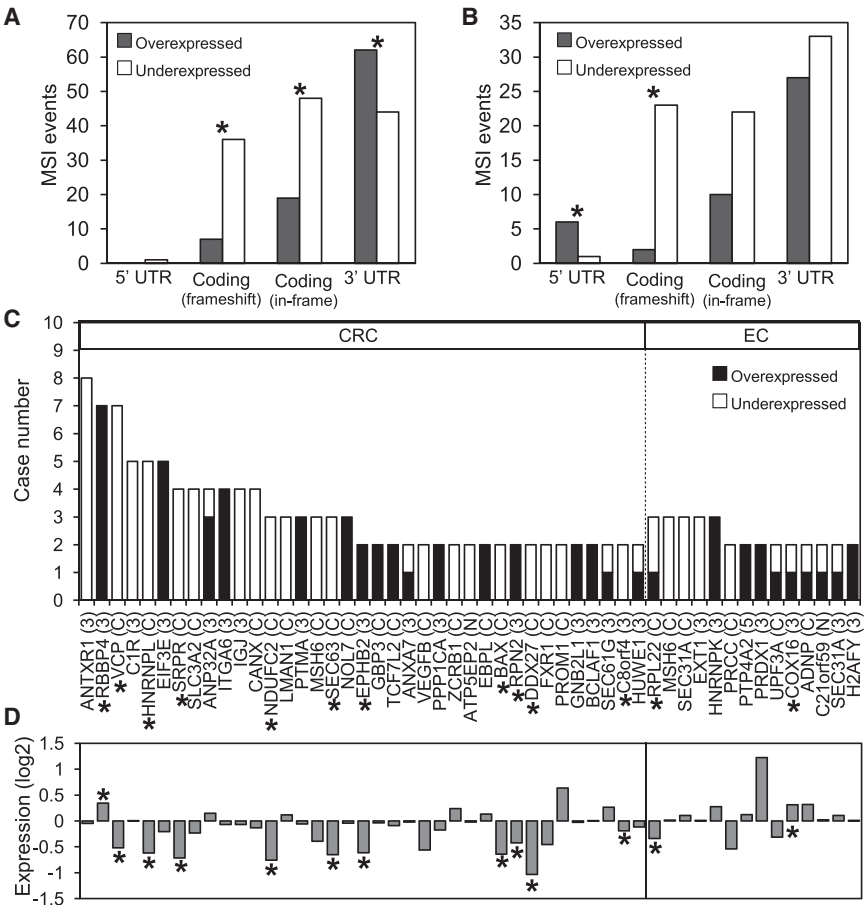

Figure 4. Association between MSI and Changes in Expression Level

(A) The MSI events in CRC genomes accompanied by a significant deviation in expression levels between the wild-type versus mutant alleles are classified into "MSI overexpressed" and "MSI underexpressed"' in each of four regions. The asterisk indicates significant differential counts (binomial test; $p < 0.05$) for frameshift coding ($p = 0.0009$), in-frame coding ($p = 0.0462$), and 3′ UTR MSI ($p = 0.0002$).

(B) Similarly for EC genomes with significant differential counts for 5′ UTR ($p = 0.0110$) and frameshift coding MSI ($p = 0.0027$).

(C) The 37 MS loci showing MSI overexpression or MSI underexpression in two or more CRC genomes are shown (x axis; left), along with 14 such MS loci from EC genomes (right). The associated gene symbols and the location of the MS (C, N, 5′, and 3′ for coding, noncoding, 5′ UTR, and 3′ UTR MSI) are shown. For each MS locus, the number of samples showing differential expression (over- or underexpressed) is plotted (y axis).

(D) The log2 ratio of the expression levels is shown (y axis). A higher ratio indicates that the gene showed higher expression in the genomes with the corresponding MSI than those without. An asterisk indicates significant (t test, $p < 0.05$) difference in the expression level.

See also Table S5.

Genome-wide Landscape of MSI

We extended our analysis to genome wide using whole-genome sequencing data from seven CRC and ten EC genomes (four and five MSI-H genomes, respectively). The number of MSI events for MSI-H genomes ranged from 11,380 to 332,565 (excluding one EC outlier with 162, which is likely to be a misclassification by the Bethesda panel), in contrast to 5 to 7,446 observed in MSS cases (Figure 5A). For subsequent analyses, we selected the six MSI-H genomes (four CRC and two EC genomes) with the largest number of MSI events. The genome-wide distribution of MS loci targeted by MSI reveals a strong depletion at coding sequences and 5′ UTRs, similar to the exome-wide mutational spectrum (Figure 5B). After normalizing for the MS counts in each category, the frequency of MSI in 3′ UTR is comparable to those in intronic or intergenic regions (Figure 5C). Analysis of MSI calls with respect to nucleotide composition and repeat length reveals high variability of mutation frequency, depending on the MS length. For instance, up to 50% and 40% of A/T and C/G mononucleotide MS with germline length 12–14 bp can have MSI in some samples, but the MSI frequency of di- and trinucleotide repeats tends to increase with longer repeats (Figure S4). Although variable genomic abundances of different MS repeat types have been reported (Subramanian et al., 2003), our results further suggest that the preference of DNA slippage events largely depends on the sequence composition and length of the repeats.

Next, we employed correlative analysis to identify genomic features associated with the occurrences of MSI. First, we find that the local MSI frequency (measured in 1 Mb bins) is inversely correlated with SNV density in four human cancer types (Figure 6A). Second, MSI frequencies are positively correlated with H3K4me3, H3K9ac, H3K36me3 and others that mark open chromatin and transcriptionally active regions but are negatively correlated with repressive histone modifications such as H3K9me2, H3K9me3, and H3K27me2 (Figure 6B). Figure S5 also shows the correlation of MSI frequency with other genomic features. The preference of DNA slippage events toward open chromatin-like domains is consistent regardless of the bin sizes used (100 kb to 10 Mb; Figure S5); when the MSI frequency across the genome was compared with the chromatin state map defined in nine human cell lines (Ernst et al., 2011), the same pattern was observed (Figure S5). Similarly, genomic segments with early, intermediate, and late DNA-replicating timing have high-to-low MSI frequencies (Figure 6C). Multiple linear regression models (Schuster-Böckler and Lehner, 2012) were adopted to examine the extent of variations in MSI frequencies that can be predicted by a combination of multiple genomic features in CRC and EC genomes (Figure S5).

The overrepresentation of cancer-specific somatic SNVs in heterochromatin-like (Schuster-Böckler and Lehner, 2012) and late-replicating domains (Koren et al., 2012) may be explained by the limited accessibility of DNA repair complexes on closed, heterochromatin-like domains (Peterson and Côté, 2004).

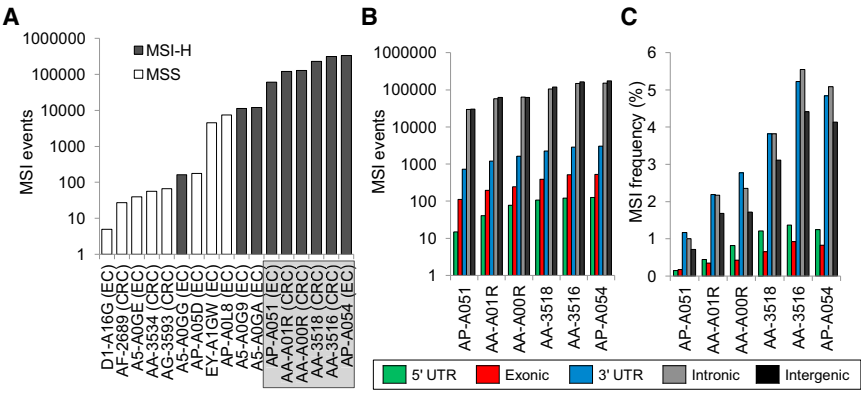

Figure 5. Genome-wide Landscape of MSI

(A) The number of MSI events genome-wide is shown for the 17 samples with whole-genome sequencing data. Six genomes (four CRC and two EC genomes) with >60,000 MSI events are shaded gray and used for subsequent analyses.

(B) The MSI events are classified into five categories based on their genomic location.

(C) The number of MSI calls is normalized by the background MS abundance in their respective regions of the genome to obtain MSI frequency. See also Figure S4.

However, this assumption is not applicable to MSI in MS-unstable genomes with a deficient MMR system. Further investigation is required to determine whether the increased MSI frequency in open chromatin-like domains arises during DNA replication or is a postreplication event. We also observed that MSI frequency is higher in introns than in intergenic regions (Figure 5C; $p = 0.002$), which is the opposite of SNV (Bass et al., 2011). The depletion of SNVs in introns is probably due to transcription-coupled repair (Pleasance et al., 2010a); elevated MSI frequency in introns suggest that MSI in MMR-deficient cancer genomes may undergo different evolutionary or fixation processes.

Finally, high-resolution analysis of the MSI frequency with respect to nucleosome occupancy demonstrated the depletion of MSI events around the positions of bulk nucleosomes as well as epigenetically modified nucleosomes H2A.Z and H3K4me3 (Figures 7A and S6). Analysis of the distances between adjacent MSI events (a pair of MSI calls separated by < 500 bp) showed two pronounced peaks at ~150 bp and ~285 bp (Figure 7B). This periodicity agrees well with the known core nucleosome size of 147 bp. Neither a depletion around nucleosomes nor a local periodicity was observed for somatic SNVs from four cancer types (Figure S6).

DISCUSSION

Our comprehensive survey of genomic loci with MSI has allowed us to gain insights on functional consequences of DNA slippage events on coding sequences and their associations with various genomic and epigenomic features. The classification of samples into the traditional MSI-H, MSI-L, and MSS categories based on the number of MSI events agreed well with the benchmark results based on the Bethesda guidelines, but the number of MSI calls was highly variable across the genomes. Besides categorizing the cancer genomes into MS-unstable and -stable ones, the number of MSI events and the related features can be useful in the evolutionary study of cancer genomes (Tsao et al., 2000).

We observed that the MSI-L and MSS categories do not show significant differences in the number of MSI events (Figure S1). Although downregulation of transcript levels or allelic loss of *MSH3* has been reported for MSI-L CRC genomes (Plaschke et al., 2012), our analysis of MSI-L and MSS genomes does not show significant differences in *MSH3* expression levels (Figure S1). Most studies of clinical correlates for CRC have observed little or no differences between MSI-L and MSS tumors and usually collapse these two groups into one (Jass, 2007). In light of our finding of similar numbers of MSI events in MSS and MSI-L tumors, we recommend discontinuation of the use of MSI-L as a distinct classification of CRC and EC tumors.

A gene-level analysis of recurrent events revealed that the ratio of frameshift-to-inframe mutations can be informative in distinguishing driver mutations from passenger ones, and both CRC and EC genomes showed a substantial level of tumor type specificity in the genes targeted by MSI. The MSI events on the TGFβ pathway genes such as *ACVR2A* and *TGFBR2* in MS-unstable CRC genomes may represent pathway-level equivalent of recurrent SNVs at other TGFβ pathway genes (e.g., *SMAD2* and *SMAD4*) in MS-stable CRC genomes (Cancer Genome Atlas Network, 2012). It was previously shown that some MS loci with recurrent MSI events in CRC genomes are not frequently altered in EC genomes (Kuismanen et al., 2002). Consistent with this, our exome-wide MSI screening clearly demonstrates tumor type specificity in recurrent MSI targets, with some novel candidates in EC genomes such as *JAK1* and *TFAM*. *JAK1* MSI may be functional given its level of recurrence (30% in MSI-H EC genomes) and its association with transcriptional downregulation of multiple gene members in the JAK-STAT pathway. The genetic perturbation of the JAK-STAT pathway was shown to decrease cellular survival of colon cancer cells in vitro (Xiong et al., 2008), which may explain the absence of the *JAK1* MSI in MS-unstable CRC genomes. Tumor-type-specific MSI targets often involve a similar molecular function, such as *MSH3* (CRC) and *PDS5B* (EC) in DNA repair processes and *AIM2* (CRC) and *TFAM* (EC) in apoptotic pathways. Elucidation of the mechanisms for tumor type-specific targeting of MSI, as well as potential molecular functions of the common and tumor type-specific mutations, will require further investigation.

Alteration of transcription levels due to an MSI event in the 3′ UTR has been attributed to the disruption of the nearby binding sites for microRNA or RNA-binding proteins (Paun et al., 2009; Yuan et al., 2009), but the impact of the MSI mutations in the coding regions and the subsequent changes on expression has not been reported previously. Our result on allele-specific expression combining transcriptome and exome sequencing data suggests that frameshift MSI events are often accompanied by lower transcript levels of the corresponding alleles. Increased frequency of SNVs in low-expressed genes has been reported

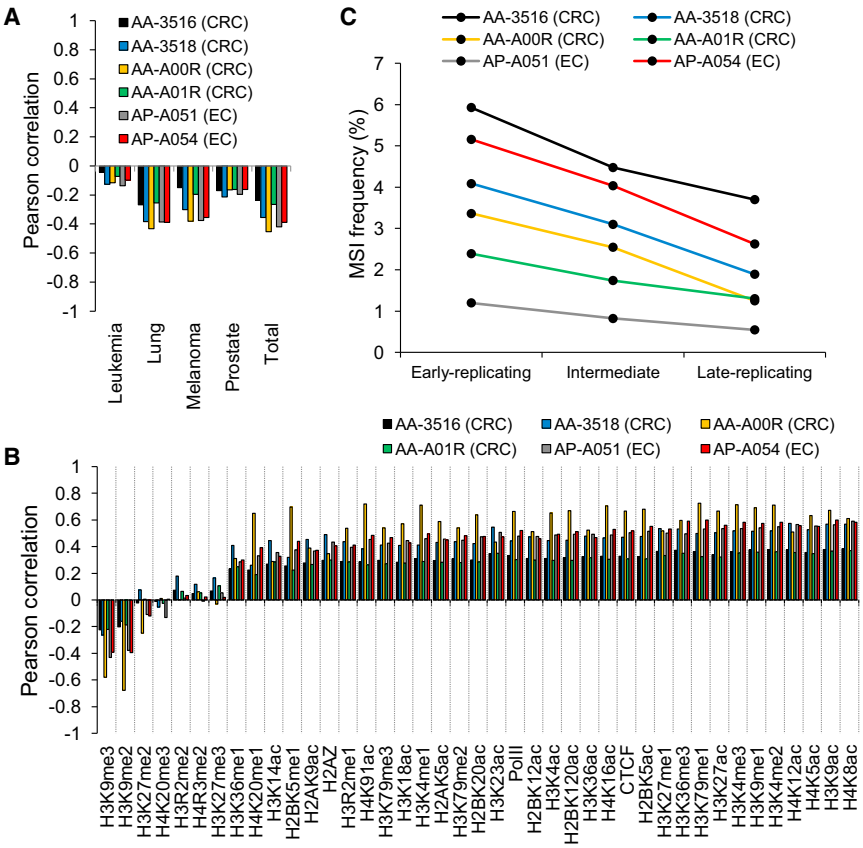

Figure 6. Correlation with Epigenomic Features

(A) The Pearson correlation between MSI frequency and SNV density (measured using 1 Mb bins) is shown for four human cancer types. For the "Total" category, SNV densities from the cancer types were combined.

(B) The same correlation analysis was performed between the frequency of MSI and enrichment of various histone modifications.

(C) MSI frequencies in the early-, intermediate-, and late-replicating timing regions are shown.

See also Figure S5.

for some cancer types (Nik-Zainal et al., 2012; Pleasance et al., 2010a, 2010b). The preference of underexpression for frameshift MSI may be associated with a known RNA surveillance pathway that eliminates mRNA containing a premature stop codon (e.g., nonsense-mediated decay) (Chang et al., 2007), which is consistent with the negative selection of the nonneutral mutations in the coding region.

In spite of the similarities between MSI and SNV, such as their preference on 3′ UTR (Pleasance et al., 2010a) and depletion on coding sequences (Bass et al., 2011), our correlative analysis revealed that their frequencies are largely anticorrelated with major differences in their regional frequencies. First, MSIs and SNVs are overrepresented in euchromatin- and heterochromatin-like domains, respectively. Second, MSIs are more enriched in introns than intergenic regions as opposed to SNVs. Third, the depletion of MSI associated with nucleosome occupancy was not observed for SNVs. For SNVs, it has been proposed that the inaccessibility of DNA repair machineries and the transcription-coupled repair are responsible for the overrepresentation of SNVs in heterochromatin-like domains and intergenic regions, respectively. However, in the context of MMR dysfunction in MS-unstable genomes, the regional preference of MSIs might have arisen during DNA replication, not as postmutational events like SNVs. One hypothesis to explain the increased MSI frequency in open chromatin is that the proofreading capabilities of DNA polymerases may be dependent on the accessibility of the chromatin. For example, the replication fork can move more rapidly in open chromatin with increased DNA slippage errors, but in closed chromatin, the slower movement of the replication fork may enhance the proofreading capabilities of DNA polymerase subunits of POLE and POLD1 (Preston et al., 2010). This chromatin-state-dependent fidelity of DNA polymerases hypothesis may also explain the decreased MSI frequencies in the nucleosome-occupied DNA segments.

The performance of our MSI-calling algorithm depends on the ability to accurately measure the length of a given MS allele. One problem with current sequencing technology is the frequent error in measuring the length of homopolymers (i.e., mononucleotide MS repeat). In this study, we used data from the Illumina platform, which uses reversible terminators that allows incorporation of just a single nucleotide at a time and is currently the most reliable platform with respect to the homopolymer issue (Dohm et al., 2008). High concordance between exome- and Sanger sequencing data for an A_{10} homopolymer (the Bethesda marker *TGFBR2*) suggests that our method performs well (Figures 2G, 2H, and S2). Illumina sequencing is still prone to a higher error rate for longer homopolymers (Minoche et al., 2011), but its impact on our analysis is minimal because ours is based on tumor versus normal comparison.

We used read alignment by Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) to associate the intraread MS repeats to the corresponding genomic loci. We have also tested additional local realignment by Genome Analysis Toolkit (GATK) (DePristo et al., 2011) or indel-sensitive Novoalign (Krawitz et al., 2010), but the number of MSI calls was very similar (Figures 2I and 2J), and the sensitivity in detecting MSI events on *TGFRB2* remained exactly the same. Local or global realignment may improve MSI calling, but a systematic evaluation will be required to delineate its platform or software dependencies.

In this study, we demonstrated that conventional exome sequencing of tumor and matched normal genomes is able to capture the exon-centric MSI events; however, there may also be some intronic and intergenic MSI events with functional significance. It was reported that MSI events near splicing sites

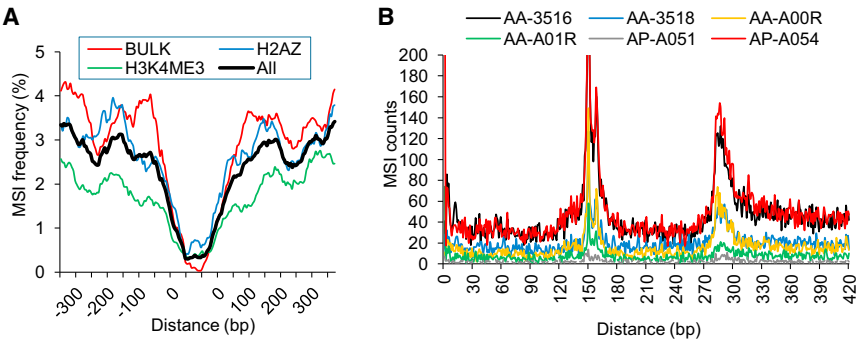

Figure 7. Depletion of MSI around Stable Nucleosome Positions

(A) MSI frequency around stable nucleosome positions is shown for one CRC genome (AA-3516; see also Figure S6).

(B) The distribution of distances between adjacent MSI pairs indicates periodicity associated with the nucleosome size.

See also Figure S6.

may alter the transcript level or splicing pattern of the target genes as shown for *MRE11* (Giannini et al., 2004) and *HSP110* (Dorard et al., 2011), respectively. In addition, the quasimonomorphic allelic nature of a Bethesda marker (BAT-26 located at the 3′ splice site of *MSH2* exon 5) in the normal population (Zhou et al., 1997) has suggested potential functional significance of MS repeats around splice sites. The availability of a larger cohort with whole-genome sequencing data will be needed to facilitate the identification of functionally important, recurrent noncoding MSI events in intronic or intergenic regions.

EXPERIMENTAL PROCEDURES

Data Sets

TCGA data were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap, accession: phs000178.v8.p7). We obtained exome data for 147 CRC and 130 EC patients, as well as whole-genome data for seven CRC and ten EC patients (tumor and matched normal genomes). All reads were 100 bp paired-end reads. We confined our analysis to those generated on the Illumina platforms.

MSI Annotation of TCGA Genomes

The MSI status (MSI-H, MSI-L, and MSS) and the clinicopathological parameters were obtained from the TCGA website (https://tcga-data.nci.nih.gov). MSI status was evaluated by TCGA using a panel of four mononucleotide repeats (BAT25, BAT26, BAT40, and TGFBRII) and three dinucleotide repeats (D2S123, D5S346, and D17S250), except for a subset of CRC genomes evaluated by five mononucleotide markers (BAT25, BAT26, NR21, NR24, and MONO27). Tumors were classified as MSI-H (>40% of markers altered), MSI-L (<40% of markers altered), and MSS (no marker altered). The methylation data of *MLH1* promoter were available for 141 CRC and 115 EC genomes.

Identification of a Reference Set of MS Repeats

To generate an exome-wide reference set of MS repeats, we downloaded the mRNA sequences of 39,496 RefSeq genes (USCS Genome Browser; hg18). We used Sputnik (http://espressosoftware.com/sputnik/) to identify MS repeats with different unit length (mono-, di-, tri-, and tetranucleotide). We limited our analysis to MS with the size 7–60 bp, as those MS could be detected accurately with the 100 bp reads, and the statistical power to detect longer repeats is lower. The frequency of MS repeats decreases logarithmically with the length of the repeats (e.g., >99% of repeats in the final set of exome and genome reference MS are smaller than 40 bp), suggesting that the vast majority of MS repeats are examined in our analysis. We found 265,862 MS in total RefSeq sequences. The repeats that encompass splice sites, have undetermined genomic coordinates, or are redundant due to multiple isoforms were removed. The filtered 146,447 MS repeats were categorized into four groups: 50,910 coding, 14,648 5′ UTR, 65,502 3′ UTR, and 15,387 noncoding (without reported coding sequences) MS, as annotated in the UCSC Genome Browser. For a genome-wide reference set of MS repeats, a total of 7,894,295 MS repeats were obtained (chromosome 1 to Y) and categorized into five groups (coding, 68,856; 5′ UTR, 15,093; 3′ UTR, 64,849; intronic, 3,193,265; intergenic, 4,552,232).

Detection of a DNA Slippage Event

The reads were aligned to NCBI build 36 (hg18) using BWA. After filtering reads with low mapping quality, intraread MS repeats were identified with the same method used to identify reference MS repeats and then intersected with the reference MS repeats by their coordinates. We required the 2 bp flanking sequences (both 5′ and 3′) of the intraread MS repeats to be identical to those of matching reference repeats, ignoring truncated MS repeats. In each genome, the distribution of the repeat allelic length at an MS locus was obtained by collecting the lengths of all intraread MS repeats mapped to that locus. We compared the distributions of MS lengths from tumor and matched normal genomes at each locus using the Kolmogorov-Smirnov statistic. An FDR of < 0.05 was used as a threshold for statistical significance, with a minimum of five tumors and five matched normal reads. We note that the number of MSI "events" refers to the absolute MSI counts per sample, whereas MSI "frequency" refers to the number of events normalized by the background MS numbers in the reference sets.

Categorization of MS Based on Allelic Length Shift

The length of MS repeat measured from each read was compared to the length of the corresponding germline MS repeat (+ and − for insertion and deletion, respectively) in the reference set. The differential read counts among different lengths were normalized to obtain relative fractions for each MSI event. Hierarchical clustering was used to group the MSI events with similar profiles. We distinguished MSI events at coding sequences into frameshift and in-frame events depending on whether the allelic length corresponding to the mode of the distribution is nontriplet or not. MSigDB v3 c5 GO categories were used for GSEA (Subramanian et al., 2005). For genes with recurrent frameshift MSI, we used the preranked version of GSEA using the level of recurrence as the weighting parameter for genes.

Allele-Specific Expression Using RNA-Seq

RNA-seq reads from MS-unstable CRC and EC tumor genomes were aligned on the RefSeq sequences using BWA. For the 1,143 and 1,224 MSI calls supported by >10 RNA-seq reads with intraread MS repeats, the differential RNA-seq read counts from wild-type and mutant alleles (depending on whether the allelic MS length is equal to that of germline or not) were compared with those from exome sequencing using Fisher's exact test. For the 37 and 14 MS loci with significant expression bias in more than one genome, the extent of differential expression was compared between the cancer genomes with or without the MSI at each locus. For gene-level expression, we used $\log_2$(RPKM + 1) from RNA-seq data (RPKM, reads per kb per million mapped reads).

Correlative Analysis with Epigenomic Features

Genome-wide features were obtained as previously described (Schuster-Böckler and Lehner, 2012). We limited the analysis to autosomal features. SNV of four human cancer types (leukemia, lung, melanoma, and prostate cancers) were downloaded from the Supplemental Data sections of the respective studies (Berger et al., 2011; Pleasance et al., 2010a, 2010b; Puente et al.,

2011). Germline polymorphisms (dbSNP build 130), GC contents, genomic coordinates of CpG islands, recombination rates, and conservation scores (all in hg18) were downloaded from the UCSC Genome Browser. Hi-C data (Lieberman-Aiden et al., 2009) were obtained from the Gene Expression Omnibus database (accession number GSE18199). For comparison with the 18 histone acetylation and 17 methylation markers, as well as the occupancy of RNA polII, CTCF, and H2AZ, the reads were downloaded as instructed in the original studies (Barski et al., 2007; Wang et al., 2008). The coordinates of the chromatin state map defined in nine human cell lines were downloaded from the UCSC genome browser. To annotate the chromatin states with respect to DNA replication timing, Repli-Seq data were obtained (Hansen et al., 2010). For GM12878, for which both chromHMM and RepliSeq data sets were available, we calculated the ratio of the early versus late replication timing (G1B and S1 versus S4 and G2) for each of the 15 chromatin states. Three chromHMM states with the highest and lowest early versus late ratio were annotated as "early" and "late" replication, with the remaining segments annotated as "Intermediate." To examine the extent of variations in MSI frequencies that can be predicted by a combination of multiple genomic features, we adopted a multiple linear regression model (Schuster-Böckler and Lehner, 2012). Fifty genomic features, including the gene expression level, were tested in an iterative manner, and the models with minimal Bayesian information criterion (BIC) were selected. The genomic occupancy profiles of nucleosomes were obtained from our previous study (Tolstorukov et al., 2011).

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.10.015.

ACKNOWLEDGMENTS

We thank The Cancer Genome Atlas Research Network for generating the data used in this work. We also thank the members of the Park laboratory (especially Drs. Eunjung Lee, Nils Gehlenborg, and Semin Lee) and Dr. Peter Kharchenko for providing comments on the manuscript, Dr. David Wheeler for helpful discussions, and the Research Information Technology Group at Harvard Medical School for providing computational resources. This work was supported by grants from the National Institutes of Health (R01 GM082798 and U24CA144025 to P.J.P.) and the National Research Foundation of Korea (2012R1A5A2047939 to T.-M.K.).

Received: January 25, 2013
Revised: July 11, 2013
Accepted: October 2, 2013
Published: November 7, 2013

REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bass, A.J., Lawrence, M.S., Brace, L.E., Ramos, A.H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A., et al. (2011). Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat. Genet. *43*, 964–968.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Boland, C.R., and Goel, A. (2010). Microsatellite instability in colorectal cancer. Gastroenterology *138*, 2073–2087, e3.

Bronner, C.E., Baker, S.M., Morrison, P.T., Warren, G., Smith, L.G., Lescoe, M.K., Kane, M., Earabino, C., Lipford, J., Lindblom, A., et al. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature *368*, 258–261.

Brough, R., Bajrami, I., Vatcheva, R., Natrajan, R., Reis-Filho, J.S., Lord, C.J., and Ashworth, A. (2012). APRIN is a cell cycle specific BRCA2-interacting protein required for genome integrity and a predictor of outcome after chemotherapy in breast cancer. EMBO J. *31*, 1160–1176.

Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Chang, Y.F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. *76*, 51–74.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. *22*, 398–406.

de la Chapelle, A., and Hampel, H. (2010). Clinical relevance of microsatellite instability in colorectal cancer. J. Clin. Oncol. *28*, 3380–3387.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105.

Dorard, C., de Thonel, A., Collura, A., Marisa, L., Svrcek, M., Lagrange, A., Jego, G., Wanherdrick, K., Joly, A.L., Buhard, O., et al. (2011). Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. Nat. Med. *17*, 1283–1289.

Duggan, B.D., Felix, J.C., Muderspach, L.I., Tourgeman, D., Zheng, J., and Shibata, D. (1994). Microsatellite instability in sporadic endometrial carcinoma. J. Natl. Cancer Inst. *86*, 1216–1221.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fernandes-Alnemri, T., Yu, J.W., Datta, P., Wu, J., and Alnemri, E.S. (2009). AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. Nature *458*, 509–513.

Giannini, G., Rinaldi, C., Ristori, E., Ambrosini, M.I., Cerignoli, F., Viel, A., Bidoli, E., Berni, S., D'Amati, G., Scambia, G., et al. (2004). Mutations of an intronic repeat induce impaired MRE11 expression in primary human cancer with microsatellite instability. Oncogene *23*, 2640–2647.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Gurin, C.C., Federici, M.G., Kang, L., and Boyd, J. (1999). Causes and consequences of microsatellite instability in endometrial carcinoma. Cancer Res. *59*, 462–466.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. USA *107*, 139–144.

Herman, J.G., Umar, A., Polyak, K., Graff, J.R., Ahuja, N., Issa, J.P., Markowitz, S., Willson, J.K., Hamilton, S.R., Kinzler, K.W., et al. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. USA *95*, 6870–6875.

Horiguchi, K., Sakamoto, K., Koinuma, D., Semba, K., Inoue, A., Inoue, S., Fujii, H., Yamaguchi, A., Miyazawa, K., Miyazono, K., and Saitoh, M. (2012). TGF-β drives epithelial-mesenchymal transition through δEF1-mediated downregulation of ESRP. Oncogene *31*, 3190–3201.

Jass, J.R. (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology *50*, 113–130.

Jung, B., Doctolero, R.T., Tajima, A., Nguyen, A.K., Keku, T., Sandler, R.S., and Carethers, J.M. (2004). Loss of activin receptor type 2 protein expression in microsatellite unstable colon cancers. Gastroenterology *126*, 654–659.

Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. *91*, 1033–1040.

Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S., and Robinson, P.N. (2010). Microindel detection in short-read sequence data. Bioinformatics *26*, 722–729.

Kuismanen, S.A., Moisio, A.L., Schweizer, P., Truninger, K., Salovaara, R., Arola, J., Butzow, R., Jiricny, J., Nyström-Lahti, M., and Peltomäki, P. (2002). Endometrial and colorectal tumors from patients with hereditary nonpolyposis colon cancer display different patterns of microsatellite instability. Am. J. Pathol. *160*, 1953–1958.

Larsson, N.G., Wang, J., Wilhelmsson, H., Oldfors, A., Rustin, P., Lewandoski, M., Barsh, G.S., and Clayton, D.A. (1998). Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. Nat. Genet. *18*, 231–236.

Leach, F.S., Nicolaides, N.C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomäki, P., Sistonen, P., Aaltonen, L.A., Nyström-Lahti, M., et al. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell *75*, 1215–1225.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W., Vogelstein, B., et al. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science *268*, 1336–1338.

Metzgar, D., Bytof, J., and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. *10*, 72–80.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. *12*, R112.

Nakayama, Y., Yamauchi, M., Minagawa, N., Torigoe, T., Izumi, H., Kohno, K., and Yamaguchi, K. (2012). Clinical significance of mitochondrial transcription factor A expression in patients with colorectal cancer. Oncol. Rep. *27*, 1325–1330.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993.

Parsons, R., Myeroff, L.L., Liu, B., Willson, J.K., Markowitz, S.D., Kinzler, K.W., and Vogelstein, B. (1995). Microsatellite instability and mutations of the transforming growth factor beta type II receptor gene in colorectal cancer. Cancer Res. *55*, 5548–5550.

Paun, B.C., Cheng, Y., Leggett, B.A., Young, J., Meltzer, S.J., and Mori, Y. (2009). Screening for microsatellite instability identifies frequent 3′-untranslated region mutation of the RB1-inducible coiled-coil 1 gene in colon tumors. PLoS ONE *4*, e7715.

Peterson, C.L., and Côté, J. (2004). Cellular machineries for chromosomal DNA repair. Genes Dev. *18*, 602–616.

Plaschke, J., Preußler, M., Ziegler, A., and Schackert, H.K. (2012). Aberrant protein expression and frequent allelic loss of MSH3 in colorectal cancer with low-level microsatellite instability. Int. J. Colorectal Dis. *27*, 911–919.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordóñez, G.R., Bignell, G.R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184–190.

Popat, S., Hubner, R., and Houlston, R.S. (2005). Systematic review of microsatellite instability and colorectal cancer prognosis. J. Clin. Oncol. *23*, 609–618.

Preston, B.D., Albertson, T.M., and Herr, A.J. (2010). DNA replication fidelity and cancer. Semin. Cancer Biol. *20*, 281–293.

Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G.R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature *475*, 101–105.

Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J.C., and Perucho, M. (1997). Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science *275*, 967–969.

Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504–507.

Subramanian, S., Mishra, R.K., and Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. *4*, R13.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tolstorukov, M.Y., Volfovsky, N., Stephens, R.M., and Park, P.J. (2011). Impact of chromatin structure on sequence variability in the human genome. Nat. Struct. Mol. Biol. *18*, 510–515.

Tsao, J.L., Yatabe, Y., Salovaara, R., Järvinen, H.J., Mecklin, J.P., Aaltonen, L.A., Tavaré, S., and Shibata, D. (2000). Genetic reconstruction of individual colorectal tumor histories. Proc. Natl. Acad. Sci. USA *97*, 1236–1241.

Umar, A., Boland, C.R., Terdiman, J.P., Syngal, S., de la Chapelle, A., Rüschoff, J., Fishel, R., Lindor, N.M., Burgart, L.J., Hamelin, R., et al. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J. Natl. Cancer Inst. *96*, 261–268.

Veigl, M.L., Kasturi, L., Olechnowicz, J., Ma, A.H., Lutterbaugh, J.D., Periyasamy, S., Li, G.M., Drummond, J., Modrich, P.L., Sedwick, W.D., and Markowitz, S.D. (1998). Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. Proc. Natl. Acad. Sci. USA *95*, 8698–8702.

Wang, J., Sun, L., Myeroff, L., Wang, X., Gentry, L.E., Yang, J., Liang, J., Zborowska, E., Markowitz, S., Willson, J.K., et al. (1995). Demonstration that mutation of the type II transforming growth factor beta receptor inactivates its tumor suppressor activity in replication error-positive colon carcinoma cells. J. Biol. Chem. *270*, 22044–22049.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Xiong, H., Zhang, Z.G., Tian, X.Q., Sun, D.F., Liang, Q.C., Zhang, Y.J., Lu, R., Chen, Y.X., and Fang, J.Y. (2008). Inhibition of JAK1, 2/STAT3 signaling induces apoptosis, cell cycle arrest, and reduces tumor cell invasion in colorectal cancer cells. Neoplasia *10*, 287–297.

Yuan, Z., Shin, J., Wilson, A., Goel, S., Ling, Y.H., Ahmed, N., Dopeso, H., Jhawer, M., Nasser, S., Montagna, C., et al. (2009). An A13 repeat within the 3′-untranslated region of epidermal growth factor receptor (EGFR) is frequently mutated in microsatellite instability colon cancers and is associated with increased EGFR expression. Cancer Res. *69*, 7811–7818.

Zhou, X.P., Hoang, J.M., Cottu, P., Thomas, G., and Hamelin, R. (1997). Allelic profiles of mononucleotide repeat microsatellites in control individuals and in colorectal tumors with and without replication errors. Oncogene *15*, 1713–1718.