Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

Tumor and Stem Cell Biology

Cancer Research

Comprehensive Evaluation of Protein Coding Mononucleotide Microsatellites in Microsatellite-Unstable Colorectal Cancer

Johanna Kondelin[1,2], Alexandra E. Gylfe[1,2], Sofie Lundgren[1,2], Tomas Tanskanen[1,2], Jiri Hamberg[1,2], Mervi Aavikko[1,2], Kimmo Palin[1,2], Heikki Ristolainen[1,2], Riku Katainen[1,2], Eevi Kaasinen[1,2], Minna Taipale[3,4], Jussi Taipale[1,2,3,4], Laura Renkonen-Sinisalo[5], Heikki Järvinen[5], Jan Böhm[6], Jukka-Pekka Mecklin[7], Pia Vahteristo[1,2], Sari Tuupanen[1,2], Lauri A. Aaltonen[1,2,3], and Esa Pitkänen[1,2]

Abstract

Approximately 15% of colorectal cancers exhibit microsatellite instability (MSI), which leads to accumulation of large numbers of small insertions and deletions (indels). Genes that provide growth advantage to cells via loss-of-function mutations in microsatellites are called MSI target genes. Several criteria to define these genes have been suggested, one of them being simple mutation frequency. Microsatellite mutation rate, however, depends on the length and nucleotide context of the microsatellite. Therefore, assessing the general impact of mismatch repair deficiency on the likelihood of mutation events is paramount when following this approach. To identify MSI target genes, we developed a statistical model for the somatic background indel mutation rate of microsatellites to assess mutation significance. Exome sequencing data of 24 MSI colorectal cancers revealed indels at 54 million mononucleotide microsatellites of three or more nucleotides in length. The top 105 microsatellites from 71 genes were further analyzed in 93 additional MSI colorectal cancers. Mutation significance and estimated clonality of mutations determined the most likely MSI target genes to be the aminoadipate-semialdehyde dehydrogenase *AASDH* and the solute transporter *SLC9A8*. Our findings offer a systematic profiling of the somatic background mutation rate in protein-coding mononucleotide microsatellites, allowing a full cataloging of the true targets of MSI in colorectal cancer. *Cancer Res; 77(15); 4078–88. ©2017 AACR.*

Introduction

Colorectal cancer is the third most common malignancy in Western countries with a mortality rate of nearly 50% (1). Approximately 15% of colorectal cancers exhibit microsatellite instability (MSI); inherited cases account for 3% whereas the remaining 12% are sporadic (2). MSI is also observed in 10%–20% of endometrial and gastric cancers (3). MSI is caused by a defect in the mismatch repair (MMR) machinery, one of the main mechanisms responsible for recognizing and repairing errors in newly synthesized DNA. MSI results from biallelic inactivation of one of the MMR genes. The most common inherited condition predisposing to colorectal cancer is Lynch syndrome where the individual carries a germline mutation in one of the MMR genes and is therefore highly predisposed to MSI colorectal cancer. In sporadic cases, the inactivation most often results from hypermethylation of the *MLH1* gene promoter (2).

MSI is characterized by accumulation of a high number of mutations across the genome, most commonly small insertions and deletions (indels) in short nucleotide tandem repeats, microsatellites. In protein coding regions of genes, these often lead to a shift in the DNA reading frame and the generation of a premature termination codon, leading to a truncated protein product. Genes that provide growth advantage to cells via loss-of-function mutations in microsatellites are called MSI target genes (4). In 1997, the National Cancer Institute workshop set criteria to distinguish MSI target genes involved in tumorigenesis from incidental mutation targets. These included (i) a high frequency of mutations, (ii) biallelic inactivation, (iii) involvement in a growth suppressor pathway, (iv) inactivation of the same pathway in MSS tumors, and (v) functional studies in *in vitro* or *in vivo* models (5). Subsequently, numerous genes have been reported as candidates for MSI target genes, many of them based on high mutation frequency alone. Only few genes have been functionally validated. Well established MSI target genes include, for example activin A receptor type 2A (*ACVR2A*) and transforming growth factor β receptor 2 (*TGFBR2*; refs. 6–8).

The tendency of a microsatellite to harbor mutations depends on the microsatellite length and the nucleotide context (9).

[1]Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland. [2]Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland. [3]Department of Biosciences and Nutrition, Karolinska Institutet, Solna, Sweden. [4]Science for Life Center, Huddinge, Sweden. [5]Department of Surgery, Helsinki University Central Hospital, Hospital District of Helsinki and Uusimaa, Helsinki, Finland. [6]Department of Pathology, Jyväskylä Central Hospital, Jyväskylä, Finland. [7]Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Esa Pitkänen, European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg 69117, Germany. Phone: 3584-0531-4252; Fax: 622-1387-8306; E-mail: esa.pitkanen@helsinki.fi

doi: 10.1158/0008-5472.CAN-17-0682

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

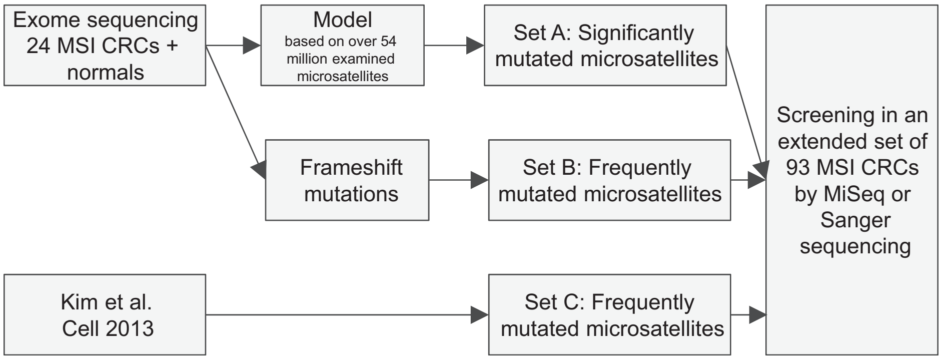

Figure 1.
Twenty-four MSI colorectal cancers and corresponding normals were exome sequenced as the discovery set. On the basis of the exome sequencing data, a statistical model was developed to account for the somatic indel mutation rates at mononucleotide microsatellites to rank genes based on mutation significance. Fifty-three most significantly mutated microsatellites from 53 genes (Set A) were selected for further validation. To compare our novel statistical model with only utilizing mutation frequencies, the most frequently mutated genes in the exome sequencing data (63 microsatellites from 35 genes; Set B) were included in the validation. Sets A and B were validated by MiSeq sequencing in 93 additional MSI colorectal cancers. To compare our results with those of another extensive NGS study, the 18 genes from the study of Kim and colleagues (18) were included in the analysis and further validated by Sanger sequencing in the 93 MSI colorectal cancers.

Therefore, when mutation frequency of microsatellites is evaluated as evidence for selection, the background mutation rate must be considered (10). Several statistical tools for calculating the somatic background mutation rate of microsatellites to identify candidate MSI target genes likely involved in tumorigenesis have been developed based on Sanger sequencing of numerous microsatellites (11–13).

In the past few years, next-generation sequencing (NGS) technologies have been widely accepted into both research and clinical use. A few large sequencing efforts on colorectal cancer and MSI have been published (14–17) but to our knowledge to date only one has focused on MSI colorectal cancer specifically (18). In that study, NGS was utilized to profile the genomic landscape of MSI in endometrial cancer and colorectal cancer. Several genes were reported to have a high mutation frequency in microsatellites. Distinguishing true driver genes from a large number of incidentally mutated passenger genes from NGS data is an obvious challenge—one that is augmented in MSI tumors that contain on average tenfold the number of mutations observed in microsatellite stable tumors (14).

As high mutation frequency is not solely sufficient to implicate a role in tumorigenesis, we designed a study to identify candidate MSI colorectal cancer target genes by statistical modeling of short indels to assess mutation significance of microsatellites. We utilized exome sequencing data from 24 sporadic MSI colorectal cancers and corresponding normals to systematically characterize the mutation profile of mononucleotide microsatellites. We observed 54,469,706 mononucleotide microsatellites with a minimum length of three nucleotides within the coding region of the genome to develop a novel statistical model for evaluation of mutation significance to discover the most likely candidate MSI target genes. As indels in microsatellites are frequent events in MSI cells, we modeled the mutation rate at mononucleotide microsatellites as the function of the nucleotide (A:T or C:G), and microsatellite length. Taking advantage of the high number of mutations detected by exome sequencing in contrast to targeted Sanger sequencing utilized in previous studies, we were able to model each microsatellite class (e.g., A:T microsatellite of eight nucleotides, A:T[8]) independently.

With the statistical model, we identified the genes with the most significantly mutated microsatellites in the exome sequencing data (Set A) with the aim of identifying the most likely MSI target genes. To compare the results from our novel statistical model to an approach utilizing mutation frequency alone, the most frequently mutated genes in our data (Set B), and a recent comprehensive NGS study on MSI colorectal cancer (Set C; ref. 18) were included in the analysis (Fig. 1).

Altogether, 105 mononucleotide microsatellites from 71 candidate MSI colorectal cancer target genes (Sets A, B, and C) were selected for further validation by MiSeq or Sanger sequencing in 93 additional MSI colorectal cancers. On the basis of mutation significance from our statistical model and the estimated clonality of mutations, we identified the most likely candidate target genes in this tumor type. Two genes, *AASDH* and *SLC9A8*, emerged as our prime candidates for novel MSI colorectal cancer target genes.

Materials and Methods

Patient material

The 24 sporadic MSI colorectal cancers and corresponding blood or normal colon tissue samples were derived from a previously characterized population-based series of 1,044 colorectal cancers for exome sequencing (Supplementary Table S1; refs. 19, 20). The study was approved by the Ethics Committee of the Hospital district of Helsinki and Uusimaa. All samples were derived either after an informed consent signed by the patient or authorization from the National Supervisory Authority for Welfare and Health. The study was conducted in accordance with Declaration of Helsinki.

Variant analysis in exome sequencing data

A novel genomic discovery tool developed in-house (BasePlayer, http://biorxiv.org/content/early/2017/04/11/126482) was used to analyze and visualize the exome sequencing data. To obtain somatic variants, the 24 tumor exomes were filtered against the 24 respective normal samples.

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

Modeling somatic indel mutations at mononucleotide microsatellites in exome sequenced samples: a novel statistical model

Indel mutation frequencies at mononucleotide microsatellites were estimated from the somatic mutation calls in exome sequencing data. Similarly to previous studies, we considered the indel mutation frequency at A:T and C:G mononucleotide microsatellites (21, 22). The indel mutation frequency was estimated separately for each nucleotide and mononucleotide microsatellite length class (e.g., A:T microsatellite of eight base pairs, or A:T[8]) for microsatellites of the human reference genome GRCh37 with a minimum length of three nucleotides. For each mutation class, we obtained the indel mutation frequency k/n, where k is the number of observed somatic indels of the particular mutation class and n is the number of callable mononucleotide microsatellite sites across the samples. A site in a sample was considered callable if the coverage at the site was at least five in both tumor and normal. Prior to analysis, positions of mutation calls, which occurred within a mononucleotide microsatellite, were adjusted to the first base of the microsatellite.

We then tested each mononucleotide microsatellite mutated in at least one tumor to discover if the site was mutated with a rate significantly higher than the somatic background mutation frequency. Here we conservatively only considered frameshift mutations in contrast to the estimation of the background mutation frequency, where all indels were considered. For each microsatellite, Fisher's exact test was used to test whether the microsatellite mutations occurred independent of the background mutation rate of the microsatellite class (e.g., A:T[8]). Specifically, the tested null hypothesis was $(m/n)/[(M-m)/(N-n)] = 1$, where m and n are the number of mutated and wild-type tumors ($m + n = 24$), and M and N are the total number of mutated and wild-type microsatellites of the same class as the tested microsatellite across the 24 tumors. The P values obtained were corrected for multiple testing with the Benjamini–Hochberg method. The microsatellites were then ranked according to their corrected P value (q-value; Set A).

MiSeq sequencing

The most significantly mutated microsatellites in the exome data, all microsatellites that were mutated with $q < 2.35 \times 10^{-4}$ (53 microsatellites from 53 genes, Set A), were selected for further validation in a set of 93 additional MSI colorectal cancers (Supplementary Table S2). To compare our approach to one where only mutation frequency is considered, the most frequently mutated genes, all genes mutated in 10 or more samples (65 microsatellites from 36 genes, Set B), were also included in the MiSeq validation (Supplementary Table S2).

Variant analysis in the MiSeq data

A comparative analysis and visualization tool developed in-house (BasePlayer) was utilized to analyze variants called in the MiSeq data. The sequencing data from the 93 tumors were filtered against whole genome sequencing data from 231 in-house normal samples of colorectal cancer patients to remove germline variants (23).

Sanger sequencing

To compare our results to those of the previously published NGS study on MSI colorectal cancer, the 18 genes published by Kim and colleagues were included in the analysis (18). In the data by Kim and colleagues there were 21 mononucleotide microsatellites in the coding region of these 18 genes that harbored frameshift mutations. In our exome data there were 10 additional coding region mononucleotide microsatellites that harbored frameshift mutations within these genes. All the 31 microsatellites in these 18 genes were selected for further validation by Sanger sequencing (Set C; Supplementary Table S2). Microsatellites in untranslated region (UTR) were excluded. The 31 microsatellites were Sanger sequenced in the set of 93 additional MSI colorectal cancers. The primer sequences and PCR conditions are shown in Supplementary Table S3.

Statistical analysis

Fisher's exact test was used to evaluate differences in mutation frequencies between experiments (exome, Sanger, MiSeq sequencing) and gene sets (Sets A, B, and C).

Results

To identify new candidates for MSI colorectal cancer target genes, and to systematically characterize mutations in mononucleotide microsatellites within the protein coding regions in this tumor type, we analyzed exome sequencing data from 24 MSI colorectal cancers and respective normal samples. We developed a novel statistical model to evaluate mutation significance in different mononucleotide microsatellite classes. The most significantly mutated mononucleotide microsatellites (53 microsatellites, from 53 genes, Set A) were screened in a targeted MiSeq sequencing of an independent set of 93 MSI colorectal cancers. To compare our novel statistical model for mutation significance to an approach where only mutation frequencies are utilized, the most frequently mutated genes (65 microsatellites from 36 genes, Set B) were also included in the MiSeq validation. To compare the results of our study to those of the previously published NGS study on MSI colorectal cancer, the 31 microsatellites from the 18 genes reported by Kim and colleagues (Set C) were further validated by Sanger sequencing in the set of 93 MSI colorectal cancers (18). The analysis workflow is summarized in Fig. 1.

Characterization of somatic microsatellite mutations in exome sequencing data of 24 MSI colorectal cancers

Analysis of the exome sequencing data of 24 MSI colorectal cancers and respective normal samples yielded a median of 2,273 somatic single nucleotide variants (SNV), 727 deletions, and 132 insertions per tumor (Fig. 2). The majority of the indels were one nucleotide deletions (51%; Supplementary Fig. S1). Exome sequencing data allowed us to investigate mutation profiles on average at 1,417,867 A:T and 851,703 C:G mononucleotide microsatellites with sufficient read coverage ($\geq$5 reads) within the coding region per tumor (Supplementary Fig. S2). Although longer C:G microsatellites are less common in the coding region than A:T microsatellites, somatic indels were found at a much higher proportion in C:G than in A:T repeats of the same length, as previously reported (Supplementary Figs. S3 and S4; ref. 11). Altogether 5,685 indels (a median of 235 per tumor) occurred at 3,955 mononucleotide microsatellites of three or more nucleotides within the coding region (GRCh37, Ensembl 71). An inverse relationship was observed between microsatellite and indel lengths, long microsatellites harboring predominantly deletions and insertions of one nucleotide (Supplementary Figs. S5 and S6). A much larger proportion of frameshifts (70%) targeted mononucleotide microsatellites than that of inframes (16%; Supplementary Fig. S7).

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

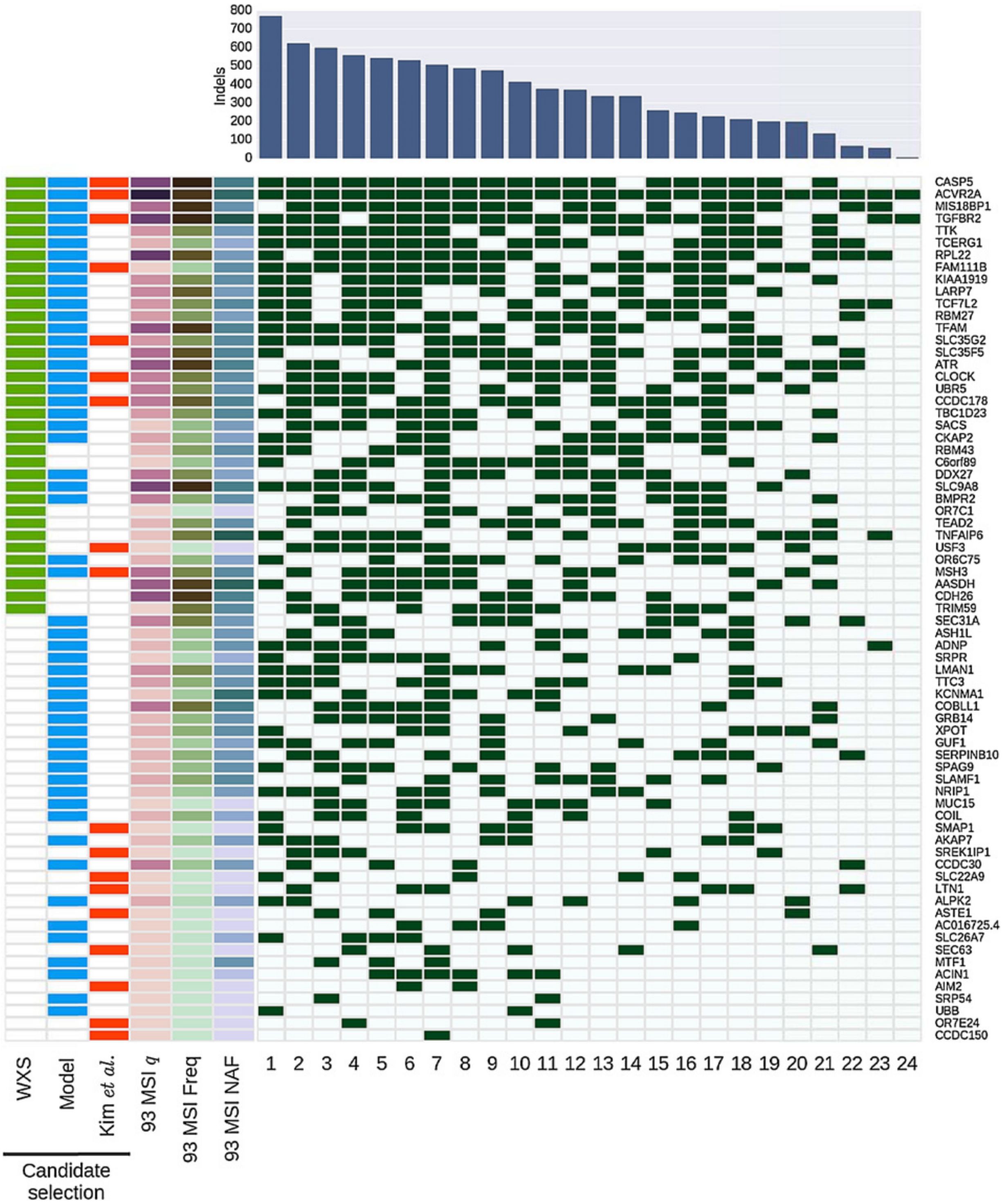

Figure 2.

A summary of the exome and targeted sequencing data of MSI tumors. Rows, the 71 candidate genes selected for validation in the extended set of 93 MSI colorectal cancers. Columns, the 24 MSI colorectal cancers subjected to exome sequencing. The genes where a tumor harbors at least one frameshift mutation in a mononucleotide microsatellite are indicated by dark green. Top, a histogram of indels within the protein coding region in 24 MSI colorectal cancers detected in exome sequencing. Left, candidate target gene selection criteria [significantly mutated according to our statistical model; frequently mutated in the whole exome sequencing (WXS) data; frequently mutated in Kim and colleagues (18)], and for the extended set of 93 MSI colorectal cancers, the statistical model *q*-value, mutation frequency (Freq), and normalized mutated allelic fraction (*NAF*). Darker color corresponds to a lower *q*-value and higher mutation frequency and allelic fraction.

Table 1. Mutation frequencies of microsatellites that were sequenced by exome sequencing, Sanger sequencing, and MiSeq

Gene	Chr	Chr position	Context	Mutation frequency in 24 MSI colorectal cancer exomes	Mutation frequency in Sanger sequencing of 93 additional MSI colorectal cancers	Mutation frequency in MiSeq sequencing of 93 additional MSI colorectal cancers	Mutation frequency by Kim et al. (18) in 27 MSI colorectal cancers
ACVR2A	2	148683686	A8	91.67	88.76	88.17	66.67
ACVR2A	2	148657041	A8	8.33	15.05	13.98	3.70
CASP5	11	104878041	T10	79.17	90.91	95.70	7.41
CASP5	11	104874011	T8	12.50	3.61	2.15	0.00
CASP5	11	104879687	T10	12.50	83.33	89.25	33.33
CASP5	11	104877976	T4	4.17	0.00	0.00	0.00
CCDC178	18	30913143	T10	45.83	68.83	70.97	29.63
CCDC178	18	30672797	T4	4.17	0.00	0.00	0.00
CCDC178	18	30928918	T4	4.17	0.00	0.00	0.00
CLOCK	4	56336954	A9	54.17	57.14	55.91	33.33
FAM111B	11	58892377	A10	70.83	65.38	74.19	33.33
MSH3	5	79970915	A8	41.67	55.81	52.69	33.33
SLC35G2	3	136573486	A9	54.17	50.00	45.16	33.33
SLC35G2	3	136573586	A7	4.17	0.00	0.00	0.00
TGFBR2	3	30691872	A10	83.33	95.45	96.77	70.37
USF3	3	113380090	T9	20.83	18.57	12.90	0.00
USF3	3	113377482	T11	20.83	96.00	78.49	59.26

Modeling short indels at mononucleotide microsatellites with a novel statistical model

Detection of signals of positive selection in MSI colorectal cancer on the basis of mutation frequency alone is biased, favoring longer mononucleotide microsatellites. To account for the effect of microsatellite length, and thus to be better able to detect driver mutations, we constructed a statistical model of indel mutation rates at mononucleotide microsatellites of variable length to rank microsatellites based on mutation significance. Microsatellites of A:T and C:G were analyzed independently of each other. In the A:T model, the accuracy of frequency estimate remained high up to the length of 16 nucleotides (<10% margin of error). However, in the C:G model similar accuracy was observed only up to eight nucleotides due to the lower number of C:G microsatellites in the coding regions (Supplementary Table S4).

We observed 3,769 mononucleotide microsatellites with at least one frameshift mutation in our data. We tested each of these sites to assess whether the site was mutated more frequently than expected considering the nucleotide and length of the microsatellite (e.g., A:T[8]). The result showed a nonuniform distribution of P values (Supplementary Fig. S8), likely stemming from unaccounted factors in this analysis such as replication timing and transcription-coupled repair (24). However, microsatellites within three previously established targets *ACVR2A*, *TGFBR2*, and *CASP5* (6, 8, 25) were among the six highest ranking genes, increasing our confidence in the statistical model (*ACVR2A*, A:T[8], mutated in 22/24 tumors, 2.12×10^{-32}, rank 1/3769; *TGFBR2*, A:T[10], 20/24, 4.13×10^{-16}, rank 4/3769; *CASP5*, A:T[10], 19/24, 1.33×10^{-14}, rank 6/3769). We selected the most significantly mutated genes, those that contained a microsatellite with $q < 2.3 \times 10^{-4}$ (53 microsatellites from 53 genes, Set A), as our primary candidates for MSI target genes were based on the exome sequencing data to be validated in the extended set of 93 MSI colorectal cancers by MiSeq sequencing (Supplementary Table S2; Supplementary Figs. S9 and S10).

Selection of candidate genes for validation based on mutation frequency alone

To compare the results from our novel statistical model to an approach where only mutation frequencies are utilized, we next identified the set of genes most frequently affected by frameshift mutations at mononucleotide microsatellites in the exome data of 24 MSI colorectal cancers. A total of 36 genes with frameshift mutations in at least 10 tumors (>41%, a mean of 14.2 frameshift mutations/gene) were found and selected for further validation in the extended set of 93 MSI colorectal cancers (65 microsatellites from 36 genes, Set B; Fig. 2; Supplementary Table S2). The three previously characterized MSI target genes, *ACVR2A*, *TGFBR2*, and *CASP5* (6, 8, 25), each containing an A:T microsatellite of at least eight nucleotides, harbored the most mutations. Indeed, of the total of 65 microsatellites occurring in the 36 genes, A:T[8] was the most common type (16/65). In addition to *ACVR2A*, *TGFBR2*, and *CASP5*, other genes in Set B that contained long microsatellites, such as *CLOCK* and *MSH3*, have been denoted MSI target genes in previous studies (26, 27).

To compare our results to those of another extensive NGS study on MSI colorectal cancer, the 18 genes reported by Kim and colleagues were included in the analysis (18). For these 18 genes, the mutation frequencies were calculated from our exome data. Also, for these 18 genes, the mutation frequency was recalculated from the mutation data of Kim and colleagues according to the same criteria that was utilized for calculating mutation frequencies from our data, the number of frameshift mutations in coding region mononucleotide microsatellites, to unify the frequencies between our data and that of Kim and colleagues. Altogether, 31 microsatellites from 18 genes (Set C) were included in further validation in the extended set of 93 MSI colorectal cancers (Supplementary Table S2).

Accurate estimation of somatic mononucleotide microsatellite mutation frequencies with targeted sequencing

The top candidate MSI target genes (the most significantly mutated genes according to our novel statistical model, Set A) and the most frequently mutated genes from our exome data (Set B) as well as from Kim and colleagues (Set C) were selected for further validation in the extended set of 93 independent MSI colorectal cancers. Altogether, this screen included (i) 53 mononucleotide microsatellites in 53 genes highlighted by our statistical model (Set A), (ii) 65 microsatellites in the 36 most frequently mutated genes in our exome sequencing data (Set B), and

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

(iii) 31 microsatellites in the 18 candidate genes of Kim and colleagues (Set C, Supplementary Table S2). As there was overlap between the sets, the entire set comprised altogether 105 microsatellites from 71 genes (Table 1; Supplementary Fig. S11). To accurately estimate the mutation frequencies at mononucleotide microsatellites, we opted for targeted sequencing of these microsatellites utilizing Illumina MiSeq or Sanger sequencing.

MiSeq sequencing and subsequent sequence analysis of 105 microsatellites in the 93 tumors yielded a median of 58.5 Mbp mapped sequencing data per tumor (Supplementary Fig. S12). The median coverage at indel sites was 184.5 reads (Supplementary Fig. S13; Supplementary Table S5). Sanger sequencing successfully amplified 30 microsatellites from 17 genes of Set C (Supplementary Table S2). One microsatellite in *OR7E24* did not amplify. The success rate of the PCR reaction varied between 100% and 60% per fragment and was on average 85% per fragment.

We then evaluated the mutation frequencies observed in the targeted NGS data. Mutation frequencies in exome and MiSeq showed good correlation (exome vs. MiSeq $\rho = 0.73$, $P = 1.29 \times 10^{-10}$; Supplementary Fig. S14). The two well-known MSI targets, *ACVR2A* and *TGFBR2* (6, 8), were mutated in almost all the tumors (Supplementary Table S2; Supplementary Fig. S14). The following 16 genes were found to have a mutation frequency >50% with 97.5% confidence: *AASDH, ACVR2A, ATR, CASP5, CCDC178, CDH26, COBLL1, USF3, LARP7, MIS18BP1, RPL22, SLC35F5, SLC9A8, TFAM, TGFBR2*, and *TRIM59* (Supplementary Table S2).

The mutation frequencies yielded by Sanger sequencing of the genes in Set C were significantly higher than those by Kim and colleagues in 14 of the 17 genes that were successfully amplified (18), likely due to the higher sensitivity of Sanger sequencing over exome sequencing (Supplementary Table S2).

Seventeen microsatellites in nine genes were included both in Set C and either Set A or Set B, and were therefore subjected to sequencing with all three methods: exome, Sanger and MiSeq sequencing (Table 1). The mutation frequencies observed in Sanger sequencing and MiSeq of the same 93 samples were close to perfect agreement (Spearman $\rho = 0.97$, $P = 2.18 \times 10^{-10}$). Although all methods showed good correlation, mutation frequencies in our exome data and the data of Kim and colleagues underestimated the frequencies relative to MiSeq and Sanger data, likely due to the higher sensitivity of the latter methods (Fig. 3; Supplementary Fig. S15).

Mutation modeling reveals novel candidate MSI colorectal cancer target genes

Analysis of mutation significance by the novel statistical model in the MiSeq sequencing data resulted in a set of 21 genes harboring a microsatellite with unexpectedly large numbers of mutations with $q < 5.99 \times 10^{-38}$ (Table 2; Supplementary Table S6). This set contained the well-established target genes *ACVR2A*

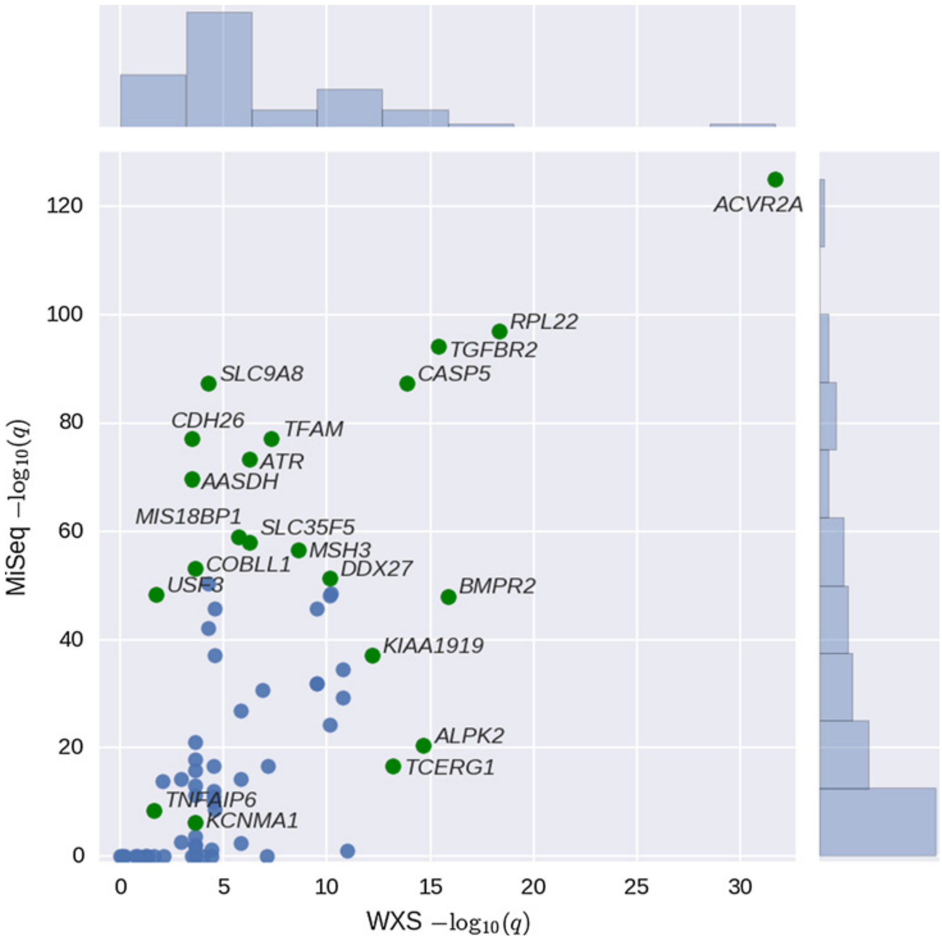

Figure 3.
Significance of the mutations in the 71 genes in the exome sequencing and targeted MiSeq data. FDR-adjusted *P* values [$\log_{10}(q)$] are shown. Genes with $\log_{10}(q) > 11$ in the exome data, the top 80% genes with respect to the *q*-value calculated in the MiSeq data ($n = 14$), and two genes highlighted in the clonality analysis (*TNFAIP6* and *KCNMA1*) are indicated (green, labeled).

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

Table 2. The top 21 genes based on mutation significance in the MiSeq data of 93 MSI colorectal cancers[a]

Gene	Chr	Chr position	Context	Sequence context	Number of mutated samples in MiSeq data of 93 MSI colorectal cancers	Mutation frequency in MiSeq data of 93 MSI colorectal cancers	*p*-value in MiSeq data of 93 MSI colorectal cancers	*q*-value in MiSeq data of 93 MSI colorectal cancers	Allelic fraction for the gene	NAF for the gene	Number of mutated samples in the exome data of 24 MSI colorectal cancers	Mutation frequency in exome data of 24 MSI colorectal cancers	*P*-value in exome data of 24 MSI colorectal cancers	*q* value in exome data of 24 MSI colorectal cancers
ACVR2A	2	148683686	A8	catAAAAAAAAgag	82	88.17	4.77E−128	6.77E−126	0.471165512	0.713763086	22	91.67	5.63E−36	2.12E−32
RPL22	1	6257785	T8	ttcTTTTTTTTgcc	70	75.27	1.28E−99	9.09E−98	0.327349973	0.385827837	16	66.67	2.60E−22	4.90E−19
TGFBR2	3	30691872	A10	aggAAAAAAAAAAgcc	90	96.77	1.48E−96	6.99E−95	0.514699856	0.820115518	20	83.33	4.39E−19	4.13E−16
CASP5	11	104878041	T10	ctgTTTTTTTTTTgtg	87	93.55	1.70E−89	4.83E−88	0.393036745	0.595087489	19	79.17	2.12E−17	1.33E−14
SLC9A8	20	48467301	T10	gggTTTTTTTTTTgtg	87	93.55	1.70E−89	4.83E−88	0.382711911	0.584909234	11	45.83	5.71E−07	5.25E−05
CASP5	11	104879687	T10	gccTTTTTTTTTTgcc	83	89.25	4.38E−81	1.04E−79	0.393036745	0.595087489	3	12.50	2.84E−01	3.29E−01
TFAM	10	60148570	A10	gacAAAAAAAAAAgtg	82	88.17	4.12E−79	7.31E−78	0.367299118	0.546325209	14	58.33	2.52E−10	4.75E−08
CDH26	20	58587784	A10	cctAAAAAAAAAAgtc	82	88.17	4.12E−79	7.31E−78	0.384991975	0.563518162	6	25.00	3.89E−09	5.86E−07
ATR	3	142274740	T10	gtaTTTTTTTTTTcag	80	86.02	2.72E−75	4.30E−74	0.375094278	0.562192296	13	54.17	3.89E−09	5.86E−07
AASDH	4	57220269	A10	cccAAAAAAAAAAtct	78	83.87	1.27E−71	1.80E−70	0.464290566	0.746116014	10	41.67	5.46E−06	3.32E−04
MIS18BP1	14	45716019	T11	caaTTTTTTTTTTTcaa	80	86.02	7.02E−61	9.07E−60	0.337040506	0.454460645	7	29.17	1.86E−02	4.90E−02
SLC35F5	2	114500277	A10	agcAAAAAAAAAAgct	71	76.34	8.90E−60	1.05E−58	0.373849711	0.539824796	13	54.17	3.89E−09	5.86E−07
MSH3	5	79970915	A8	gacAAAAAAAAggg	49	52.69	2.67E−58	2.91E−57	0.373316768	0.500575462	10	41.67	1.19E−11	2.35E−09
COBLL1	2	165551296	A9	tgcAAAAAAAAAgag	57	61.29	6.56E−55	6.65E−54	0.408966512	0.579573128	8	33.33	3.21E−06	2.28E−04
DDX27	20	47858504	A8	gccAAAAAAAAggg	46	49.46	4.17E−53	3.95E−52	0.303034983	0.328848044	11	45.83	2.97E−13	7.48E−11
CCDC178	18	30913143	T10	ccaTTTTTTTTTTata	66	70.97	4.15E−52	3.69E−51	0.378491814	0.543255425	11	45.83	5.71E−07	5.25E−05
CCDC30	1	43002197	A4	gtcAAAAgag	16	17.20	3.55E−50	2.97E−49	0.322159348	0.378154201	4	16.67	1.95E−13	6.11E−11
BMPR2	2	203420130	A7	gggAAAAAAAccg	32	34.41	5.39E−50	4.25E−49	0.369959838	0.46641945	11	45.83	1.15E−19	1.44E−16
UBR5	8	103289349	T8	ttcTTTTTTTTgcc	44	47.31	9.77E−50	7.30E−49	0.351124147	0.452660683	11	45.83	2.97E−13	7.48E−11
USF3	3	113377482	T11	gggTTTTTTTTTTTagc	73	78.49	1.16E−49	8.26E−49	0.435918295	0.687129175	5	20.83	1.49E−01	1.75E−01
CLOCK	4	56336954	A9	gctAAAAAAAAAcca	52	55.91	2.99E−47	1.93E−46	0.369834671	0.478032211	13	54.17	1.45E−12	3.03E−10
SEC31A	4	83785565	T9	caaTTTTTTTTTggc	52	55.91	2.99E−47	1.93E−46	0.336380704	0.401873511	9	37.50	2.40E−07	2.82E−05

[a] Summary of the 21 genes containing the most significantly mutated microsatellites based on the significance analysis of the targeted MiSeq data; mutation frequency in the exome and MiSeq data, mutation significance in the exome and MiSeq data, NAF of the genes

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

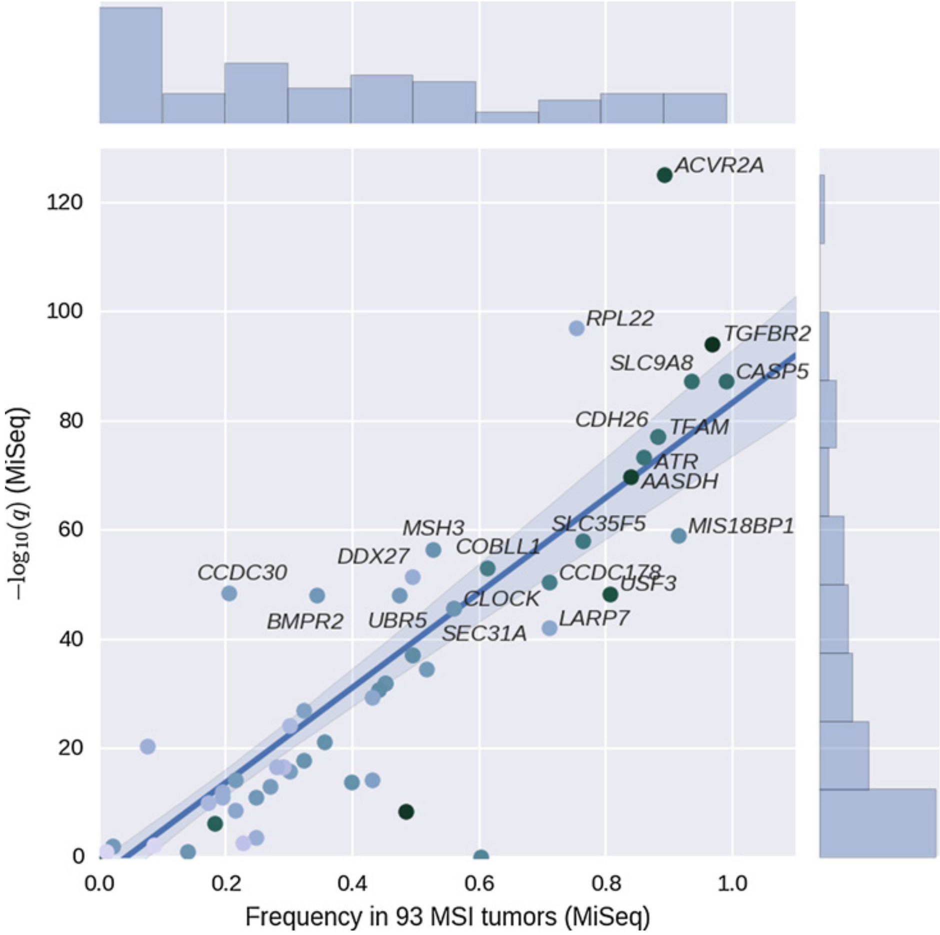

Figure 4.
Mutation frequency and significance [$-\log_{10}(q)$] in the MiSeq data shown for the 71 genes selected for screening (Sets A, B, and C) in the extended set of 93 MSI colorectal cancers. Color indicates the normalized allelic fractions (*NAF*) of the gene, with darker shades corresponding to higher *NAF* values. For each gene, the smallest *q* value of the repeat sites within the gene is shown. Genes with $-\log_{10}(q) > 40$ are labeled. Regression line with 95% confidence intervals is shown.

(ranked 1st) and *TGFBR2* (3rd; refs. 6, 8), as well as *AASDH*, *ATR*, *BMPR2*, *CASP5*, *CCDC178*, *CCDC30*, *CDH26*, *CLOCK*, *COBLL1*, *DDX27*, *USF3*, *MIS18BP1*, *MSH3*, *RPL22*, *SEC31A*, *SLC35F5*, *SLC9A8*, *TFAM*, and *UBR5* (Fig. 4; Supplementary Table S7). In subsequent Sanger sequencing, the *ALPK2* mutations were found to be germline polymorphisms.

To gain insight into the clonality of mutations and thus additional evidence for driver genes, we examined NAFs of indels in the 71 genes sequenced with MiSeq (Supplementary Fig. S16). As expected, mutations in the known MSI targets *TGFBR2* and *ACVR2A* (6, 8) were found to be highly clonal (mNAFs 88% and 81%, respectively; Supplementary Fig. S17; Supplementary Table S6). However, indels in *CASP5* displayed much lower allelic fractions (mNAF = 67%) as compared to *TGFBR2* and *ACVR2A*, suggesting that not all *CASP5* mutations are early events in tumor development. A total of eleven genes displayed a normalized allelic fraction of at least 60% (Supplementary Fig. S16). Indels in *TNFAIP6* appeared particularly clonal (mNAF = 89%); in subsequent analysis this gene was found to harbor a large amount of additional germline variation in mononucleotide microsatellites. Moreover, many of the somatic indels in *TNFAIP6* were inframe mutations, reducing the likelihood that this gene is a true driver via inactivation of the gene. Besides the well-characterized MSI target genes, both *AASDH* and *SLC9A8* were found to be frequently affected by relatively clonal indels (*AASDH*, 78/93 mutated tumors, mNAF = 83%; *SLC9A8*, 87/93 mutated tumors, mNAF = 66%). In addition, both *AASDH* and *SLC9A8* were highlighted in the significance analysis of the MiSeq data (*q*-value ranks; *AASDH* 10/142, *SLC9A8* 5/142; Fig. 5; Supplementary Table S2). Finally, *KCNMA1* was mutated only in 17 of 93 tumors but displayed a relatively high normalized allelic fraction (mNAF = 72%), appearing as an outlier in this analysis (Supplementary Fig. S16).

In conclusion, among the most significant and the most clonal genes we found the previously established target genes *ACVR2A* and *TGFBR2* (6–8), but also genes such as *AASDH* and *SLC9A8* that to our knowledge have not been associated with MSI cancers before (Fig. 5; Table 2; Supplementary Table S6).

Discussion

Comprehensive understanding of mutations contributing to human cancers is essential for improved diagnostics and new therapeutic approaches and personalized care. The challenge of distinguishing driver mutations from passengers is augmented in MSI tumors due to their high mutation load. In regard to identifying driver genes, the mutation frequency of a gene alone is a poor predictor of causality, and additional parameters should be considered (28).

To identify new target genes in MSI colorectal cancer, we utilized a discovery set of 24 MSI colorectal cancers to gain insight into the patterns of small indels in different mononucleotide

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

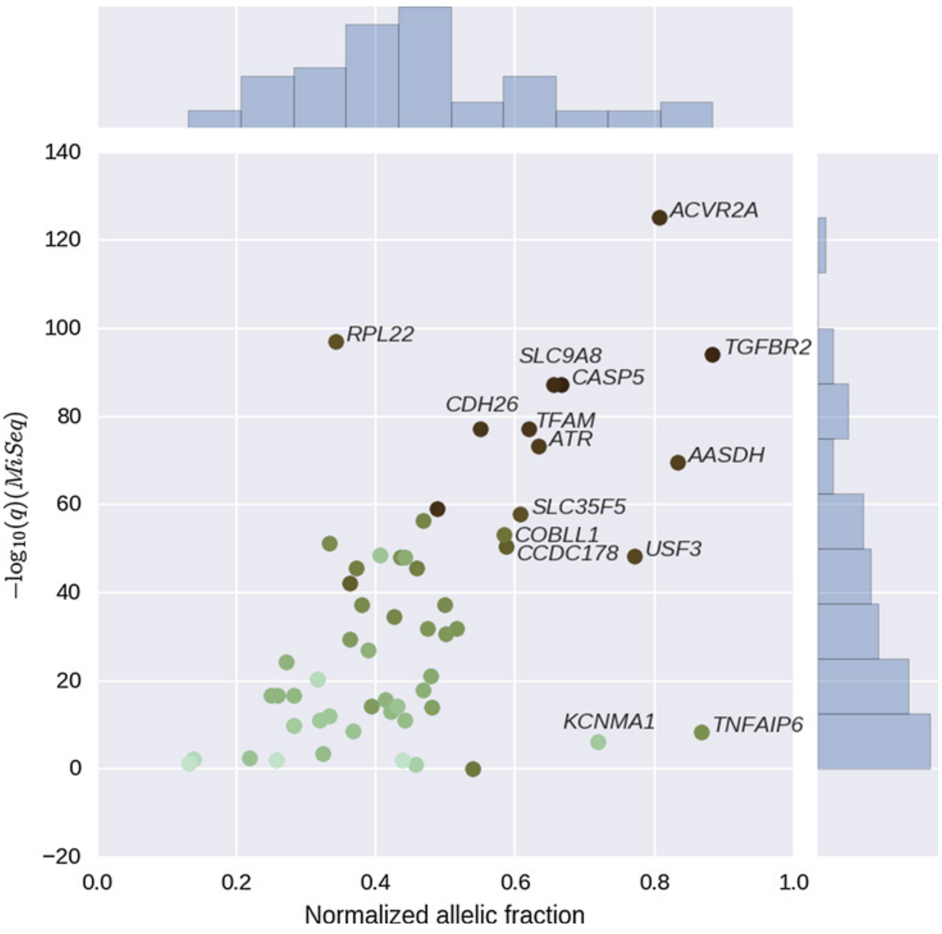

Figure 5.
Mean normalized allelic fraction (mNAF) and mutation significance in the MiSeq data shown for the 71 genes selected for screening (Sets A, B, and C) in the extended set. Color indicates the mutation frequency of the gene, with darker shades corresponding to higher frequencies. Genes that have $q < 1 \times 10^{-60}$ or mNAF > 0.55 are labeled.

microsatellite classes. We developed a novel statistical model to rank genes based on mutation significance of microsatellites and consequently identified genes harboring microsatellites with unexpectedly high numbers of frameshift mutations in relation to the length and type of the mononucleotide microsatellite (Supplementary Table S2). From this list we selected 53 microsatellites from 53 genes (Set A) to be further validated by MiSeq sequencing in an additional set of 93 MSI colorectal cancers. To compare the results of our statistical model to utilizing only mutation frequencies, the most frequently mutated genes in the exome sequencing data (Set B, 65 microsatellites from 36 genes) were included in the MiSeq screening.

Analysis of mutation significance in the MiSeq data by the statistical model resulted in a ranked list of candidate genes (Fig. 4 and Supplementary Table S6). Enabled by the high coverage of the MiSeq data, we contrasted this ranking with the estimated clonality of the mutations to gain additional evidence for driver mutations (Fig. 5; Supplementary Table S6). This set of genes serves as our candidates for MSI colorectal cancer target genes. Among the most significant and the most clonal genes were the previously established target genes *ACVR2A* and *TGFBR2* (6–8), but also genes such as *AASDH* and *SLC9A8* that to our knowledge have not been associated with MSI cancers before (Fig. 5; Table 2; Supplementary Table S6). These two genes are the prime novel candidates for MSI colorectal cancer target genes emerging from this analysis.

AASDH (aminoadipate-semialdehyde dehydrogenase) encodes an enzyme that has been suggested to activate β-alanine (29). It is located in a region of chromosome 4 that exhibits copy number loss in early onset colorectal cancer (30). A deletion of 2Ts in a 3T-microsatellite has been reported in this gene in familial colorectal cancer of unknown cause (31). Exposure of MCF-7 breast cancer cells on β-alanine resulted in reduced migration and proliferation of the cells as well as in increased sensitivity to doxorubicin (32). The frameshift mutations observed in our data suggest an inactivating effect on the protein product. Inactivation of AASDH could result in reduced activation of β-alanine and hence increased migration and proliferation of cells as well as decreased sensitivity to doxorubicin. MSI tumors have been shown to exhibit resistance to doxorubicin treatment (33–34). The mechanism of doxorubicin resistance has been linked to several transporter proteins (35), yet to our knowledge the mechanism of resistance related to MSI remains elusive. MSI colorectal cancers have also been shown to be resistant to 5-fluorouracil (5-FU; ref. 2). The main metabolite of 5-FU is α-fluoro-β-alanine (36). Inactivation of AASDH could possibly affect the processing of this metabolite.

SLC9A8 (solute carrier family 9 member A8) encodes a transmembrane protein that exchanges extracellular Na^+ for intracellular H^+ (37). It has been reported to be involved in intestinal mucosal integrity by regulating the functions of goblet and Paneth cells (38), and its loss has been shown to result in reduced mucin

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

production and increased bacterial adhesion (39). The expression of *SLC9A8* is inhibited by TNFα (39) and EGF (40).

From the top 21 genes with the most significantly mutated microsatellites in the MiSeq data, nine (*ACVR2A*, *ATR*, *BMPR2*, *CASP5*, *CLOCK*, *MSH3*, *SLC35F5*, *TGFBR2*, and *TFAM*) were previously described MSI target genes for which functional data to support their role in tumorigenesis has been reported (Supplementary Tables S6 and S7; Supplementary Fig. S18). From four genes (*CCDC178*, *RPL22*, *SEC31A*, and *UBR5*) mutation data have been reported (11, 18, 41). Two of the top genes (*MIS18BP1* and *USF3*) have been reported in MSI gastric cancer (42). Six of the top genes (*AASDH*, *CCDC30*, *CDH26*, *COBLL1*, *DDX27*, and *SLC9A8*) have to our knowledge not been reported in MSI cancers before.

Detection of signals of positive selection in MSI colorectal cancer on the basis of mutation frequency alone biases detection toward longer mononucleotide microsatellites. To compare the results from our statistical model to utilizing only mutation frequency of the gene, the most frequently mutated genes in the exome data (Set B, 65 microsatellites from 36 genes) were included in the MiSeq screening (Supplementary Table S2). From the top 21 most frequently mutated genes in the exome sequencing data, only 12 (*ACVR2A*, *ATR*, *CASP5*, *CCDC178*, *CLOCK*, *DDX27*, *MIS18BP1*, *RPL22*, *SLC35F5*, *TFAM*, *TGFBR2*, and *UBR5*) were found among the top 21 genes containing the most significantly mutated microsatellites in the MiSeq data (Supplementary Tables S2 and S6).

Microsatellites have extensively been studied for a few decades and the importance of accounting for the effect of MMR on the mutation spectrum has been stressed in various efforts (11, 13, 21). Yet many studies overlook this phenomenon and new candidate genes are continuously published based on mutation frequency alone (43–45). To our knowledge, only one other comprehensive NGS study focusing on MSI colorectal cancer has been published to date (18). In our study we identified the same previously well-established MSI target genes (*ACVR2A*, *TGFBR2*, and *CASP5*) as Kim and colleagues. However, in our study we identified 61 genes with a mutation frequency over 30%, whereas in the study by Kim and colleagues 18 such genes were reported (Supplementary Table S2; Supplementary Fig. S19). Of the 18 genes of Kim and colleagues, only seven (*ACVR2A*, *CASP5*, *CCDC178*, *CLOCK*, *USF3*, *MSH3*, and *TGFBR2*) were identified among the 21 genes containing the most significantly mutated microsatellites in our MiSeq data (Fig. 2; Supplementary Table S6).

To date, a few statistical models based on targeted sequencing of a set of microsatellites have been utilized in study of somatic mutations in microsatellites (11, 13, 21). Woerner and colleagues implemented a nonlinear regression model to identify candidate driver genes in MSI colorectal cancer, and from this analysis nine candidate genes emerged (13). Six of these (*ACVR2A*, *TGFBR2*, *TCF7L2*, *MSH3*, *BAX*, and *ASTE1*) were also found to be mutated in our data (Supplementary Table S2). To our knowledge, however, our study represents the first ever effort to systematically characterize the mutation landscape of coding region mononucleotide microsatellites in relation to the expected somatic background mutation frequency to identify MSI target genes.

In investigation of somatic mutations in microsatellites, considering the impact of MMR deficiency on the accumulation of mutations is fundamental. When comparing our results to those of another comprehensive study (18), and to other previous studies, the importance of accounting for the background mutation rate in evaluation of mutation frequencies in microsatellites is highlighted. Our novel statistical model provides a new reference for the expected somatic mutation rate in mononucleotide microsatellites and thus a novel tool for analyzing mutation significance in MSI colorectal cancers. Utilizing a discovery set larger than that in our study might enable identification of more novel candidates for MSI colorectal cancer target genes. In addition to MSI colorectal cancer, our approach should be considered when studying other MSI cancers. Finally, 25 years after the original discovery of MSI (2), our statistical model enables the construction of a comprehensive catalogue of the candidate main target genes in MSI tumors.

Disclosure of Potential Conflicts of Interest

A.E. Gylfe is R&D Lead at Human Longevity, Inc. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: J. Kondelin, P. Vahteristo, L.A. Aaltonen, E. Pitkänen
Development of methodology: J. Kondelin, R. Katainen, L.A. Aaltonen, E. Pitkänen
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): L. Renkonen-Sinisalo, H. Järvinen, J. Böhm, J.-P. Mecklin, P. Vahteristo, S. Tuupanen
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J. Kondelin, A.E. Gylfe, T. Tanskanen, J. Hamberg, K. Palin, H. Ristolainen, E. Kaasinen, S. Tuupanen, L.A. Aaltonen, E. Pitkänen
Writing, review, and/or revision of the manuscript: J. Kondelin, A.E. Gylfe, T. Tanskanen, M. Aavikko, R. Katainen, E. Kaasinen, J. Taipale, L. Renkonen-Sinisalo, J. Böhm, P. Vahteristo, L.A. Aaltonen, E. Pitkänen
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S. Lundgren, J. Hamberg, R. Katainen, M. Taipale, L.A. Aaltonen
Study supervision: J. Taipale, L.A. Aaltonen, E. Pitkänen

Acknowledgments

The authors thank Heikki Metsola, Sini Nieminen, Sirpa Soisalo, Marjo Rajalaakso, Inga-Lill Svedberg, Iina Vuoristo, and Alison Ollikainen for technical assistance. We acknowledge the computational resources provided by the ELIXIR node, hosted at the CSC–IT Center for Science, Finland. Dr. Tae-Min Kim and Professor Peter Park are thanked for kindly providing us their mutation data.

Grant Support

L.A. Aaltonen received funding from the Academy of Finland (Finnish Center of Excellence Program 2012–2017; 1250345); The Finnish Cancer Society, Sigrid Juselius Foundation, Jane and Aatos and Erkko Foundation, and SYSCOL (an EU FP7 Collaborative Project), 258236. J.K. Kondelin received funding from Biomedicum Helsinki Foundation, Otto Malm Foundation, Ida Montin Foundation, Orion-Farmos Research Foundation, Oskar Öflunds Stiftelse, Maud Kuistila Memorial Foundation, and The Doctoral Programme in Biomedicine in the Doctoral School of Health Sciences at University of Helsinki. K.J. Palin received funding from the Nordic Information for Action eScience Center (NIASC), the Nordic Center of Excellence financed by NordForsk (project 62721).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 13, 2017; revised May 24, 2017; accepted June 5, 2017; published OnlineFirst June 13, 2017.

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61:69–90.
2. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology 2010;138:2073–87.
3. Hamelin R, Chalastanis A, Colas C, El Bchiri J, Mercier D, Schreurs AS, et al. Clinical and molecular consequences of microsatellite instability in human cancers. Bull Cancer 2008;95:121–32.
4. Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 2002;62:2447–54.
5. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, et al. A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248–57.
6. Jung B, Doctolero RT, Tajima A, Nguyen AK, Keku T, Sandler RS, et al. Loss of activin receptor type 2 protein expression in microsatellite unstable colon cancers. Gastroenterology 2004;126:654–9.
7. Jung BH, Beck SE, Cabral J, Chau E, Cabrera BL, Fiorino A, et al. Activin type 2 receptor restoration in MSI-H colon cancer suppresses growth and enhances migration with activin. Gastroenterology 2007; 132:633–44.
8. Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 1995;268:1336–8.
9. Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol 1997;17:2851–8.
10. Zhang L, Yu J, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B. Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. Cancer Res 2001;61:3801–5.
11. Alhopuro P, Sammalkorpi H, Niittymaki I, Bistrom M, Raitila A, Saharinen J, et al. Candidate driver genes in microsatellite-unstable colorectal cancer. Int J Cancer 2012;130:1558–66.
12. Duval A, Rolland S, Compoint A, Tubacher E, Iacopetta B, Thomas G, et al. Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. Hum Mol Genet 2001;10:513–8.
13. Woerner SM, Benner A, Sutter C, Schiller M, Yuan YP, Keller G, et al. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative real common target genes. Oncogene 2003;22:2226–35.
14. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
15. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature 2012;488:660–4.
16. Lin EI, Tseng LH, Gocke CD, Reil S, Le DT, Azad NS, et al. Mutational profiling of colorectal cancers with microsatellite instability. Oncotarget 2015;6:42334–44.
17. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 2016;22:1342–50.
18. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 2013;155:858–68.
19. Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomaki P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481–7.
20. Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol 2000;18:2193–200.
21. Sammalkorpi H, Alhopuro P, Lehtonen R, Tuimala J, Mecklin JP, Jarvinen HJ, et al. Background mutation frequency in microsatellite-unstable colorectal cancer. Cancer Res 2007;67:5691–8.
22. Vilkki S, Launonen V, Karhu A, Sistonen P, Vastrik I, Aaltonen LA. Screening for microsatellite instability target genes in colorectal cancers. J Med Genet 2002;39:785–9.
23. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet 2015;47:818–21.
24. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
25. Schwartz S Jr, Yamamoto H, Navarro M, Maestro M, Reventos J, Perucho M. Frameshift mutations at mononucleotide repeats in caspase-5 and other target genes in endometrial and gastrointestinal cancer of the microsatellite mutator phenotype. Cancer Res 1999;59:2995–3002.
26. Alhopuro P, Bjorklund M, Sammalkorpi H, Turunen M, Tuupanen S, Bistrom M, et al. Mutations in the circadian gene CLOCK in colorectal cancer. Mol Cancer Res 2010;8:952–60.
27. Edelmann W, Umar A, Yang K, Heyer J, Kucherlapati M, Lia M, et al. The DNA mismatch repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression. Cancer Res 2000;60:803–7.
28. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
29. Drozak J, Veiga-da-Cunha M, Kadziolka B, Van Schaftingen E. Vertebrate acyl CoA synthetase family member 4 (ACSF4-U26) is a beta-alanine-activating enzyme homologous to bacterial non-ribosomal peptide synthetase. FEBS J 2014;281:1585–97.
30. Berg M, Agesen TH, Thiis-Evensen E, INFAC-study group, Merok MA, Teixeira MR, et al. Distinct high resolution genome profiles of early onset and late onset colorectal cancer integrated with gene expression data identify candidate susceptibility loci. Mol Cancer 2010;9:100. doi: 10.1186/1476-4598-9-100.
31. DeRycke MS, Gunawardena SR, Middha S, Asmann YW, Schaid DJ, McDonnell SK, et al. Identification of novel variants in colorectal cancer families by high-throughput exome sequencing. Cancer Epidemiol Biomarkers Prev 2013;22:1239–51.
32. Vaughan RA, Gannon NP, Garcia-Smith R, Licon-Munoz Y, Barberena MA, Bisoffi M, et al. Beta-alanine suppresses malignant breast epithelial cell aggressiveness through alterations in metabolism and cellular acidity in vitro. Mol Cancer 2014;13:14.
33. Claij N, te Riele H. Microsatellite instability in human cancer: a prognostic marker for chemotherapy? Exp Cell Res 1999;246:1–10.
34. Fedier A, Schwarz VA, Walt H, Carpini RD, Haller U, Fink D. Resistance to topoisomerase poisons due to loss of DNA mismatch repair. Int J Cancer 2001;93:571–6.
35. Thorn CF, Oshiro C, Marsh S, Hernandez-Boussard T, McLeod H, Klein TE, et al. Doxorubicin pathways: pharmacodynamics and adverse effects. Pharmacogenet Genomics 2011;21:440–6.
36. Rubino FM, Verduci C, Buratti M, Fustinoni S, Campo L, Omodeo-Sale E, et al. Assay of urinary alpha-fluoro-beta-alanine by gas chromatography-mass spectrometry for the biological monitoring of occupational exposure to 5-fluorouracil in oncology nurses and pharmacy technicians. Biomed Chromatogr 2006;20:257–66.
37. Xu H, Chen H, Dong J, Lynch R, Ghishan FK. Gastrointestinal distribution and kinetic characterization of the sodium-hydrogen exchanger isoform 8 (NHE8). Cell Physiol Biochem 2008;21:109–16.
38. Wang A, Li J, Zhao Y, Johansson ME, Xu H, Ghishan FK. Loss of NHE8 expression impairs intestinal mucosal integrity. Am J Physiol Gastrointest Liver Physiol 2015;309:G855–64.
39. Xu H, Li Q, Zhao Y, Li J, Ghishan FK. Intestinal NHE8 is highly expressed in goblet cells and its expression is subject to TNF-alpha regulation. Am J Physiol Gastrointest Liver Physiol 2016;310:G64–9.
40. Xu H, Zhang B, Li J, Chen H, Tooley J, Ghishan FK. Epidermal growth factor inhibits intestinal NHE8 expression via reducing its basal transcription. Am J Physiol Cell Physiol 2010;299:C51–7.
41. Ferreira AM, Tuominen I, van Dijk-Bos K, Sanjabi B, van der Sluis T, van der Zee AG, et al. High frequency of RPL22 mutations in microsatellite-unstable colorectal and endometrial tumors. Hum Mutat 2014;35:1442–5.
42. Yoon K, Lee S, Han TS, Moon SY, Yun SM, Kong SH, et al. Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in korean gastric cancers. Genome Res 2013;23: 1109–17.
43. Jo YS, Kim MS, Yoo NJ, Lee SH. Frameshift mutations of AKAP9 gene in gastric and colorectal cancers with high microsatellite instability. Pathol Oncol Res 2016;22:587–92.
44. Lee JH, Song SY, Kim MS, Yoo NJ, Lee SH. Frameshift mutations of a tumor suppressor gene ZNF292 in gastric and colorectal cancers with high microsatellite instability. APMIS 2016;124:556–60.
45. An CH, Je EM, Yoo NJ, Lee SH. Frameshift mutations of cadherin genes DCHS2, CDH10 and CDH24 genes in gastric and colorectal cancers with high microsatellite instability. Pathol Oncol Res 2015;21:181–5.

Published OnlineFirst June 13, 2017; DOI: 10.1158/0008-5472.CAN-17-0682

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

American Association for Cancer Research

Comprehensive Evaluation of Protein Coding Mononucleotide Microsatellites in Microsatellite-Unstable Colorectal Cancer

Johanna Kondelin, Alexandra E. Gylfe, Sofie Lundgren, et al.

Cancer Res 2017;77:4078-4088. Published OnlineFirst June 13, 2017.