A Single-Nucleotide Substitution Mutator Phenotype Revealed by Exome Sequencing of Human Colon Adenomas


Abstract

Oncogene-induced DNA replication stress is thought to drive genomic instability in cancer. In particular, replication stress can explain the high prevalence of focal genomic deletions mapping within very large genes in human tumors. However, the origin of single-nucleotide substitutions (SNS) in nonfamilial cancers is strongly debated. Some argue that cancers have a mutator phenotype, whereas others argue that the normal DNA replication error rates are sufficient to explain the number of observed SNSs. Here, we sequenced the exomes of 24, mostly precancerous, colon polyps. Analysis of the sequences revealed mutations in the APC, CTNNB1, and BRAF genes as the presumptive cancer-initiating events and many passenger SNSs. We used the number of SNSs in the various lesions to calculate mutation rates for normal colon and adenomas and found that colon adenomas exhibit a mutator phenotype. Interestingly, the SNSs in the adenomas mapped more often than expected within very large genes, where focal deletions in response to DNA replication stress also map. We propose that single-stranded DNA generated in response to oncogene-induced replication stress compromises the repair of deaminated cytosines and other damaged bases, leading to the observed SNS mutator phenotype. Cancer Res; 72(23); 6279–89. ©2012 AACR.

Introduction

Human cancers are characterized by the presence of genomic instability (1–4). One form of genomic instability, chromosomal instability (CIN), is present in most cancers and refers to changes in chromosome number and structure. Multiple mechanisms can lead to CIN in cancers (1–3). A specific type of aberrant chromosome structure, focal chromosomal deletions, may be caused by oncogene-induced DNA replication stress (4–6). An analysis of thousands of human cancers and hundreds of cancer cell lines has revealed that focal deletions are observed at high frequency within specific chromosomal sites known as common fragile sites (CFS; refs. 7, 8). Deletions within CFSs are pathognomonic of DNA replication stress. In cells exposed to chemicals that inhibit DNA replication, such as aphidicolin and hydroxyurea, genomic deletions occur predominantly within CFSs (9–12). Furthermore, in various experimental models, activated oncogenes induce DNA replication stress and deletions within CFSs that are very similar to the deletions induced by aphidicolin and hydroxyurea (13–15). Finally, DNA replication stress is prevalent in human precancerous lesions and cancers and is associated with focal deletions within CFSs (16–18). Taken together, these observations suggest that oncogene-induced DNA replication stress is responsible for the focal deletions and LOH observed at CFSs in human precancerous lesions and cancers (4–6).

Another form of genomic instability, microsatellite instability (MSI), is characterized by changes in the number of short-nucleotide repeats, called microsatellites (1–3). MSI is observed only in a minority of cancers and is caused by mutation or epigenetic inactivation of DNA mismatch repair genes (1–3).

Finally, it has been argued that cancers have yet another form of genomic instability, which makes them susceptible to acquiring single-nucleotide substitutions (SNS). This type of genomic instability is often referred to as the mutator phenotype (19–21). The mutator phenotype hypothesis posits that a gene important for maintaining genomic stability is inactivated early in cancer development, presumably in precancerous lesions or even in normal cells, and that the subsequent increase in mutation rate allows the acquisition of cancer-driver mutations. Defects in DNA repair genes have been proposed as the most likely culprits for the mutator phenotype (19–21).
Despite the elegant arguments in favor of the existence of a mutator phenotype, it should be noted that an increased SNS mutation rate has not yet been experimentally shown in precancerous lesions. Accordingly, the mutator phenotype hypothesis has been heavily debated (22–25). The proponents argue that a mutator phenotype is the only way to explain how a cell can acquire the multitude of mutations needed for malignant transformation. The opponents have argued that mutations in DNA repair genes have not been identified in the majority of nonhereditary human cancers and, especially, not in precancerous lesions (23). Furthermore, that normal DNA replication error rates are sufficient to explain the number of SNSs observed by genome-wide sequencing of human cancers (24).

It should be stated that, in broad terms, the mutator phenotype hypothesis embraces all types of genomic instability, including CIN and MSI (19–21). However, the recent debate has used a more narrow definition, focusing entirely on SNSs, as the presence of CIN and MSI in cancers is widely accepted (24, 25). Hence, here, we will also use the narrow definition of mutator phenotype, referring only to SNSs.

High-throughput DNA sequencing is a powerful method to analyze the presence of SNSs in human clinical samples and, thus, to help address the mutator phenotype debate. Thus, we explored whether exome-wide DNA sequencing of precancerous lesions could provide new insights on the mechanisms underlying genomic instability, as well as on the sequence of events leading to cancer development. Other than pancreatic cysts (26, 27), no human precancerous lesions have been subjected to exome sequencing to our knowledge.

Materials and Methods

Clinical and histopathologic data
Macroscopically appearing adenomas and patient-matched normal colorectal tissue were endoscopically removed at the Laiko General Hospital (Athens, Greece) after approval of the study by the Ethics Committee of the University of Athens Medical School (Athens, Greece) and obtaining informed consent from the patients. All patients were asymptomatic and were examined in the context of regularly scheduled colonoscopies. Only the patient with the adenocarcinoma case CA2 was subjected to colonoscopy because of complaint of blood in the stools. The patients were not related to each other and did not have a family history of predisposition to colon cancer or other cancers, as assessed by their medical histories. Relevant clinicopathologic data are shown in Supplementary Table S1. Resected polyps and normal colon tissue from the same patient were snap-frozen and stored at −70°C. Histologic diagnosis was conducted by 2 pathologists (V.G. Gorgoulis and I.S. Pateras), who evaluated the samples independently of each other, using formalin-fixed and paraffin-embedded sections. Immunohistochemistry was conducted in paraffin-embedded tissues, using the UltraVision LP Detection System (Thermo Scientific) and an antibody specific for phosphorylated histone H2AX (p-Ser139, 05-636, Millipore).

DNA extraction and exome sequencing
The overall methodology was as previously described (28). DNA was extracted from the frozen tissues using the QIAamp DNA Mini Kit (Qiagen), exome capture was conducted using the SureSelect Human Exon v3 50Mb (Agilent Technologies) or SeqCap EZ Human Exome Library SR version 1.2 35 Mb (Roche-Nimblegen) kits, and sequencing was conducted on Illumina HiSeq2000 and GAIIx instruments with paired-end 105 nt reads. Burrows–Wheeler Aligner (BWA) software was used to align the sequence reads to the human reference genome NCBI build (GRCh37/hg19). SAMtools was used to remove PCR duplicates and to call single-nucleotide variants (SNV). Detection of small insertions and deletions (smINDEL) was conducted with Pindel 0.2.2 software. The search for somatic mutations was restricted to the regions that were covered at least 10-fold in both the normal and polyp tissue samples. The average sequencing coverage was 155× for polyp DNA and 146× for normal tissue DNA (Supplementary Table S1).

Calling of SNSs
The initial list of SNVs was filtered against the common (>5%) germline polymorphisms present in the db135 and 1,000 genomes databases. SNVs present in the normal tissue sample from the same patient at a frequency of greater than 2% were also filtered out. To call somatic mutations, we improved, as described later, upon the algorithms that we had used for our melanoma exome sequencing project (28). These improvements enhanced the accuracy of calling SNSs (Supplementary Fig. S1). In contrast to SNVs, smINDELs were called with lower accuracy and, therefore, we report only those smINDELS that were validated by Sanger sequencing. For both SNVs and smINDELS, we focused on the mutations that map to the protein coding sequences and to splice sites, as the untranslated exonic regions were less well covered by the commercially available exome capture kits.

To call SNSs, a modified SAMtools score was calculated for each SNV using the following formula:

\[
\text{Modified SAMtools score} = \frac{|\log(TNR \times 100/AC)|^2 \times \log(IS)}{\text{BAL}^2 + 1},
\]

where TNR is the total number of reads spanning a SNV, AC is the exome-wide average coverage for the specific tissue, IS is the initial SAMtools score, BAL equals NFRW/(NFRW + NRRW) − NFRV/(NFRV + NRRV), where NFRW is the number of forward-direction reads for the wild-type sequence, NRRW is the number of reverse direction reads for the wild-type sequence, NFRV is the number of forward-direction reads for the variant sequence, and NRRV is the number of reverse direction reads for the variant sequence.

The above pipeline was repeated twice, switching the labels of the polyp and normal tissue datasets. The modified SAMtools scores from the switched datasets were used to establish a threshold for calling SNSs. The threshold was the mean score plus 3 SDs (Supplementary Fig. S1).

Two hundred and sixty SNVs were subjected to validation by Sanger sequencing. Included within this list were 15 SNVs that were less than the threshold for calling SNSs, as a means to test the SNS-calling method described earlier. Of the 260 SNVs...
subjected to Sanger sequencing, 241 were confirmed. The final number of called SNSs was 1,147 after very few adjustments were made to incorporate the results of the validation experiments. Potential cancer-driver mutations were evaluated using the MuSiC algorithm (29).

**LOH and copy number alterations**

Two independent methods were used to detect regions of LOH and copy number alterations (CNA) using the exome-sequencing data. Both methods relied on single-nucleotide polymorphisms (SNP) that were heterozygous in the normal tissue DNA and on the total number of reads across the genome. In the first method, ratios of SNPs in the normal versus polyp tissue and median values of number of sequencing reads were averaged over 10 Mb bins and visualized using CIRCOS plots. The second method used an updated version of a previously described pipeline (30). Reads were aligned to the hg19 reference genome and only those that aligned to regions targeted by exome sequence capture probes were used in subsequent steps. When the regions targeted by 2 or more capture probes overlapped, all consecutive overlapping regions were first merged and subsequently split into bins of 200 bp. A copy number ratio was calculated separately for each bin. Between sample differences in sequencing coverage were normalized by transforming read counts into RPKM (reads per kilobase of bin length per million mapped reads) values for each bin.

Mutation frequencies

A total of 19,075 protein-coding RefSeq genes were separated into bins according to their genomic size. The expected frequency of mutations in each bin was calculated on the basis of the fraction of coding sequence length (CDS) of the genes in each bin relative to the total CDS of all genes or based on the fraction of genomic length of the genes in each bin relative to the total genomic length of all genes, depending on the data (coding sequences or genome-wide sequences) being analyzed. Differences in the observed and expected frequencies were evaluated using the cumulative binomial distribution function.

**Results**

**Cancer-driver mutations**

The colon is an ideal system to study human precancerous lesions (31). In this tissue, cancer proceeds via a classic sequence of hyperplasia, dysplasia, and cancer (Fig. 1A). Typically, the precancerous and cancerous lesions protrude into the lumen of the colon as polyps, which can be resected, providing material for DNA sequencing. For this study, we sequenced the exomes of 24 randomly selected polyps obtained from 22 individuals with no family history of predisposition to cancer. DNA from normal tissue from all these patients was also sequenced. Histologic analysis revealed that one of these polyps was a hyperplastic polyp (no dysplasia), 17 were adenomas of the tubulovillous type (with mild, moderate, or severe dysplasia), 3 were serrated adenomas, and 3 were adenocarcinomas (Fig. 1A and Supplementary Table S1). For all polyps, the exome sequencing data generated a list of SNSs and smINDELs. In addition, regions of genomic CNAs and LOH were identified by analyzing the number of sequencing reads corresponding to polymorphic alleles.

To explore the pathways by which normal colon cells become transformed, we tabulated the SNSs and smINDELs targeting the genes that have been implicated in colon cancer development (31–34). Mutations in the APC, CTNNB1, and BRAF genes were found to be mutually exclusive in our series, allowing us to subdivide the polyps into 3 distinct groups (Fig. 1B).

The group with APC mutations included 20 polyps, of which 1 was a hyperplastic polyp, 16 were adenomas (all of the tubulovillous type), and 3 were adenocarcinomas. Of the 30 APC mutations identified in the 20 polyps, 16 were smINDELs and 14 were SNSs with the majority of polyps having at least 1 smINDEL (Fig. 1B). All the APC mutations introduced premature termination codons that, in many cases, mapped within residues 1,304 to 1,435 of the APC protein, consistent with the spectrum of APC mutations previously reported in sporadic colorectal carcinomas (31–34). Furthermore, none of the polyps had retained a wild-type APC allele; either 2 distinct APC mutations were identified per polyp or a single mutation was associated with LOH (Fig. 1B and Supplementary Figs. S2 and S3). The only exceptions were the hyperplastic polyp and an adenoma with moderate dysplasia (ML3), in which only a single APC mutation was identified. The hyperplastic polyp is histologically a very early lesion with no signs of dysplasia; whereas, the adenoma ML3 had an inactivating mutation in the FBXW11/beta-TrCP2 gene. The protein product of FBXW11 functions together with APC and FAM123B/WTX to target beta-catenin for degradation (35). Interestingly, mutations targeting the FAM123B gene were also detected in 2 adenomas (MI3 and SE1) in our series (Fig. 1B).

The group with CTNNB1 mutations included only 1 polyp, an adenoma with moderate dysplasia, and tubulovillous histology (Fig. 1B). The mutation in this polyp was homozygous, due to LOH (Supplementary Fig. S3), and resulted in a Ser45Phe substitution within beta-catenin, the protein product of CTNNB1. This substitution prevents beta-catenin from being degraded and, therefore, is equivalent to APC inactivation (35).

The group with BRAF mutations included the 3 adenomas with serrated histology (Fig. 1B), consistent with previous reports linking activated BRAF to this adenoma type (36). All 3 serrated adenomas in our study had a recurrent point mutation that resulted in a Val600Glu substitution and renders the B-Raf kinase constitutively active (31, 36). Unlike the APC and CTNNB1 mutations, all the BRAF mutations in our series were heterozygous (Supplementary Fig. S3).

In the adenomas with CTNNB1 and BRAF mutations, no mutations in other known cancer-driver genes were identified by exome sequencing. However, in the adenomas with APC mutations we identified additional cancer-driver mutations, whose number correlated with the degree of dysplasia and invasiveness (Fig. 2A and Supplementary Table S2). Specifically, mutations were identified in the KRAS, NRAS, GNAS,
AKT1, ARID1A, SOX9, and TP53 genes. All these mutations were SNSs, with the exception of 2 smINDEL mutations targeting SOX9. The high prevalence of SNSs contrasts with the spectrum of mutations targeting APC, more than half of which were smINDELs.

KRAS-activating mutations targeting Gly12 or Gly13 were identified in an adenoma with moderate dysplasia (MI4), 2 adenomas with severe dysplasia (SE3 and SE4), and 1 adenocarcinoma (CA3). Another adenocarcinoma (CA2) had an NRAS-activating mutation targeting Gly12. Finally, 1 adenoma with mild dysplasia (MI4) also had a KRAS mutation, but this resulted in an Ala146Thr substitution, which has been reported in human cancers but at a greatly reduced frequency compared with the substitutions targeting Gly12 or Gly13 (37).

A mutation in ARID1A was identified in an adenoma with moderate dysplasia (MS2); ARID1A regulates chromatin structure and is very frequently inactivated in colorectal carcinomas and other human cancers (38). SOX9, which, like ARID1A, also regulates chromatin structure and transcription (39), was mutated in 2 adenomas (MI3 and SE1) and 1 adenocarcinoma (CA1). Finally, 1 adenoma with moderate dysplasia (ML2), 1 with severe dysplasia (SE2), and 1 adenocarcinoma (CA2) had inactivating mutations in TP53, the gene that encodes the p53 tumor suppressor (Fig. 2A).

In addition to SNSs, the exome sequencing also identified CNAs and regions of LOH. This analysis revealed that mutated oncogenes, such as KRAS, often mapped to genomic regions that had LOH, but no CNA, making the mutant allele homozygous; whereas, the TP53 gene often mapped to regions that exhibited LOH, which was occasionally accompanied by a focal genomic deletion (Fig. 2A and Supplementary Fig. S3). In 1 adenoma with severe dysplasia (SE3), we identified a focal amplification of the MYC gene (Fig. 2A and Supplementary Fig. S4), whereas 2 adenomas and 2 of 3 adenocarcinomas had LOH targeting the q arm of chromosome 18, in which the SMAD4 gene maps (Fig. 2A and Supplementary Fig. S3).
The list of cancer-driver events described earlier is clearly not exhaustive. Several other SNSs seemed interesting and need to be investigated further. For example, mutations in PCDH10, a gene that encodes a cell adhesion molecule and that is considered to function as a tumor suppressor in gastric and esophageal cancer (40), was mutated in 3 polyps (MI4, SE1, and CA1; Fig. 2A), which is statistically significant (Supplementary Table S2). It is also interesting that other genes that are frequently mutated in human colon adenocarcinomas (31–34), such as PIK3CA and FBXW7, were not found to be mutated in our adenoma series, suggesting that mutations in these genes are acquired at later stages of colon cancer development. The full list of identified SNSs, the majority of which seem to be passenger mutations, is presented in Supplementary Table S3.

Distinct events in the evolution of adenomas
Like others (41), we reasoned that the frequencies of the mutant alleles (cancer-driver and passenger SNSs) might provide insights about the order in which specific mutations were acquired during cancer progression. Indeed, in an adenoma with mild dysplasia (MH), the frequencies of the mutant alleles mapping to diploid regions of the genome (in blue) were distributed within 2 major peaks, suggesting the presence of a subclone, whose expansion was driven by a recent cancer-driver event (Fig. 2B). The APC SNS mapped to the peak with the highest mutant allele frequency, consistent with it being an early event. Interestingly, the frequencies of the mutant alleles that mapped to the parts of the genome with LOH (in yellow), were distributed in 2 clearly distinct peaks (Fig. 2B), reminiscent of the distribution of polymorphic alleles in LOH regions (Supplementary Fig. S2). From this profile, we conclude that the LOH event affecting chromosome 12, in which KRAS maps, was more recent than the oncogenic KRAS SNS. This example, also illustrates that SNSs that had been acquired early during cancer development may have low allele frequencies, if they have been targeted by more recent LOH events.

The distribution plots of the mutant alleles in a small adenoma with moderate dysplasia (MS3) also revealed several peaks of allele frequencies (Fig. 2B). Interestingly, the 2 mutant APC alleles had distinct frequencies suggesting that they were acquired at different times in the development of this adenoma. In a large adenoma with moderate dysplasia (ML4), recent LOH events were evident, whereas the low frequency of a mutant KRAS allele in an adenoma with severe dysplasia (SE3) suggested that the KRAS mutation was a recent event (Fig. 2B). In contrast to the adenomas, the frequency distributions of the mutant alleles in the adenocarcinomas (CA1 and CA2) were quite complex (Fig. 2B), making it difficult to draw conclusions about the natural history of these lesions.
Mutation rates in human adenomas

The information obtained from exome sequencing of precancerous lesions may have use beyond identifying cancer-driver events. An important question, that is currently debated, is whether precancerous lesions have a mutator phenotype (19–25). Opponents of the mutator phenotype hypothesis argue that normal DNA replication error rates can account for 60,000 SNSs per cell in fast replicating tissues, such as the colon, by middle age (24). This translates to 350 SNSs per exome, assuming that mutation rates are 2-fold lower in highly expressed genes, as compared with the rest of the genome (42). The proponents of the mutator phenotype argue, instead, for lower basal DNA replication error rates and for acquisition of point mutations primarily after transformation of the normal tissue to a precancerous lesion (21, 25).

In an attempt to address this debate, we plotted the number of SNSs observed in the precancerous lesions as a function of the size of the polyp and the age of the patient (Fig. 3A). The number of SNSs significantly correlated ($P < 0.02$) with polyp size and, importantly, the regression line intercepted the x- and y-axes at their origins (0 mutations at size of 0 cm), suggesting that the majority of SNSs were acquired after oncogenic transformation (Fig. 3A). Furthermore, the number of SNSs did not correlate to patient age in a statistically significant manner ($P = 0.67$), as would have been predicted by the opponents of the mutator phenotype hypothesis (Fig. 3A). Indeed, the number of SNSs observed in the various polyps suggests that the opponents of the mutator phenotype hypothesis have overestimated the DNA replication error rate in normal tissues, as some polyps had less than 10 SNSs, instead of the 350 SNSs that had been predicted (24).

To better document the presence of a mutator phenotype in colon polyps, we determined estimates of mutation rates for both normal colon tissue and colon adenomas. Mutation rates are expressed as the number of acquired mutations per base pair per cell division. To explain our method, we first consider a theoretical example (Fig. 3B–D). Before the acquisition of the first cancer-driver mutation (mutation D1), a cell in an adult organism may have accumulated somatic mutations, which will be referred to as passenger mutations P0. Because the polyp originates from a single cell, the P0 mutations would be present in all cells of the polyp (Fig. 3B). Assuming 50% normal tissue contamination in the sequenced DNA sample and heterozygous P0 mutations, the P0 mutant allele frequencies would be about 25%. The frequency of the mutant D1 allele would also be in this range (Fig. 3C). Passenger mutations that are acquired after the D1 mutation (mutations P1.1, P1.2, etc.; Fig. 3B) would be present in a subset of the cells of the

![Figure 3](cancerres.aacrjournals.org/). Evidence for a mutator phenotype in colon adenomas. A, correlation of the number of SNSs to polyp size (diameter) and patient age for all precancerous lesions in our study (hyperplastic polyp, tubulovillous, and serrated adenomas). The level of statistical significance is indicated and, when significant, the regression line is also indicated. B–D, theoretical model showing changes in mutant allele frequencies during tumor progression. B, accumulation of passenger (P) and driver (D) mutations in a polyp. P0, passenger mutation acquired before the first cancer-driver (D1) mutation; P1.1 and P1.2, distinct passenger mutations acquired in different cells after acquisition of mutation D1; D2, second driver mutation. C, mutant allele frequencies in a polyp with a single driver mutation (mutation D1) and 3 passenger mutations (P0, P1.1, and P1.2). Alleles with frequencies less than 10% are not called by current high-throughput sequencing algorithms. D, mutant allele frequencies in the same polyp after acquisition of a second driver mutation. E, calculated mutation rates for normal colon tissue and for adenomas with multiple cancer-driver events (see Supplementary Fig. S5 for more details).
polyp and would not be identified by high-throughput sequencing, as their mutant allele frequencies would typically be less than 10% threshold value of the mutation-calling algorithm (Fig. 3C). Therefore, the number of SNSs observed in polyps with a single cancer-driver event can be used to calculate normal tissue mutation rates. These polyps, which are the hyperplastic polyp (HP1) and the 3 serrated adenomas, had on average 5 ± 3 (mean ± 1 SD) P0 mutations per polyp (Fig. 1B and Supplementary Fig. S5).

In the same theoretical example, a second cancer-driver mutation (mutation D2; Fig. 3B), which confers a growth advantage to a specific cell and its progeny, will allow detection of a specific subset of P1 mutations, as their allele frequencies exceed the 10% threshold (Fig. 3D). Therefore, in polyps with more than 1 cancer-driver events, high-throughput sequencing can identify passenger mutations acquired after oncogenic transformation, allowing mutation rates for the precancerous lesions to be calculated. In our sample, the tubulovillous adenomas, all of which had more than 1 cancer-driver event, had 50 ± 28 (mean ± 1 SD) SNSs per polyp (Fig. 1B and Supplementary Fig. S5).

In addition to the number of SNSs, we also need an estimate of the number of cell divisions over which the specific mutations were acquired. The crypt stem cells of the colon divide every 3 to 4 days (45), which means that a normal colon stem cell of a 60-year-old individual would have undergone about 6,000 cell divisions since birth, which is exactly the same estimate as used by others (44). With 5 ± 3 (mean ± 1 SD) P0 mutations per polyp, the normal colon stem cell mutation rate can be calculated as 1.2 ± 0.7 × 10⁻¹¹ SNSs per base pair per cell division (Fig. 3E and Supplementary Fig. S5), a number similar to the mutation rate observed in the testis (45, 46), which in adult males is also a continuously dividing tissue.

For the tubulovillous adenomas, we estimate, based on clinical experience (31), that they develop on an average over a 5-year period. Tritiated thymidine labeling studies further establish that in cancer cells, certain loci are more prone to genomic instability than others. For example, 7 of the 10 most common focal deletions in human cancers map to very large genes (7, 8). These deletions have been attributed to oncogene-induced DNA replication stress (6), on the basis of experiments showing that genes greater than 0.6 Mb in size are particularly prone to genomic deletions, when cells are treated with chemicals that induce DNA replication stress (10–12) or when oncogenes are overexpressed (13–15). Indeed, most of the so-called CFSs, which are genomic loci sensitive to DNA replication stress, map within very large genes (48–50).

To examine whether the SNSs were also preferentially distributed within very large genes, we grouped the mutations according to the size of the gene that they targeted and then calculated observed and expected mutation frequencies (Fig. 4B). In the 4 adenomas with mild dysplasia, the genes with genomic size greater than 0.6 Mb were targeted by SNSs 3.5 times more frequently than expected (P < 6.9 × 10⁻¹⁶). This analysis took into account CDS length (Supplementary Table S4). Adenomas with a moderate degree of dysplasia showed a 3-fold increase in the frequency of SNSs within very large genes, which dropped to 2-fold for the adenocarcinomas (Fig. 4B). When the SNSs from all adenomas and adenocarcinomas with APC mutations were included (19 polyps; 1,097 SNSs), the statistical significance reached a level of P < 5.2 × 10⁻⁶ (Fig. 4B).

To determine whether the higher prevalence rates of SNSs in very large genes could be reproduced in other datasets, we examined the mutations recently reported in 224 human colorectal cancer exomes (34). This dataset does not include adenomas but could still be used to validate the 2-fold effect that we observed in the adenocarcinomas (Fig. 4B). The 224 colorectal cancers were divided by the authors of the study into 2 groups: an MSI-positive and an MSI-negative group. In the MSI-negative group, the number of mutations in the very large genes was 2-fold higher than expected, whereas in the MSI-positive group, the effect of gene size on mutation frequency was much smaller (Fig. 4C). The difference in behavior of the MSI-negative and MSI-positive cancers was statistically significant at a level of P < 6.9 × 10⁻¹⁶. We subsequently examined datasets of genome-wide mutations in colorectal cancers (33) and datasets of exomic mutations present in head and neck cancers (51) and in melanomas (52). In all these datasets, a higher prevalence of SNSs was observed within the very large genes (Fig. 4C). As with the analysis of the exomic data, the analysis of the genome-wide data took gene size into account (Supplementary Table S5).

Discussion

The exome sequencing of colon precancerous lesions helps improve our understanding of cancer development. First, we propose that mutations in the APC, CTNNB1, or BRAF genes
represent the cancer-initiating event. This assumption is based on the known biology of these genes (31, 35) and the fact that all polyps had a mutation in one of these genes. Thus, it seems that only a limited set of events can initiate oncogenic transformation in the colon: activation of the Wnt signaling pathway or the B-Raf kinase. In a recent study of MCF10A cells grown as mammospheres, only the **ERBB2** oncogene, of many tested, could initiate aberrant growth (53). We speculate that similar constraints operate in the human colon, explaining why, for example, we did not observe **KRAS** mutations in the absence of **APC** mutations. Furthermore, we note that certain genes that are frequently mutated in colon cancers (31–34), such as **PIK3CA** and **FBXW7**, were not found to be mutated in our adenoma series, suggesting that the mutations targeting these 2 genes are later events.

![Figure 4](cancerres.aacrjournals.org)

**Figure 4.** Mutation spectra of SNSs and preferential targeting of very large genes. **A**, mutation spectra of SNSs. The SNSs were classified according to the mutation (C to A, C to G, C to T, T to A, T to C, and T to G) and according to the nucleotides at the 5' and 3' positions relative to the mutated (M) nucleotide. The polyps are abbreviated as shown in Fig. 1 and the number of polyps included in the analysis is indicated in parentheses. **APC** TVA, tubulovillous adenomas with mild, moderate, or severe dysplasia bearing **APC** mutations. The asterisk indicates the activating **BRAF** mutation in the serrated adenomas. **B**, ratios of observed over expected number of SNSs mapping to genes of the indicated genomic sizes. The polyps are abbreviated as shown in Fig. 1 and the number of polyps included in the analysis is indicated in parentheses. In each graph, the number of mutations mapping to genes greater than 0.6 Mb in size and the total number of mutations is indicated (separated by /), followed by the statistical probability \( P \) for rejecting the null hypothesis for the bin of genes with genomic size greater than 0.6 Mb. All \( P \) values shown are statistically significant. **APC** TVA + **CA**, all the tubulovillous adenomas with **APC** mutations and the 3 adenocarcinomas. The cumulative coding length of the genes greater than 0.6 Mb in size corresponds to 1.90% of the cumulative coding length of the 19,075 protein coding genes used in our analysis (Supplementary Table S4). **C**, ratios of observed over expected number of SNSs mapping to genes of the indicated genomic sizes in published datasets. From top to bottom: exome dataset of 224 colorectal cancers comprising 187 MSI-negative cancers and 31 MSI-positive (with a high degree of MSI) cancers (34); genome-wide dataset of 9 colorectal cancers (33); exome dataset of 74 head and neck squamous cell carcinomas (51); and exome dataset of 121 melanomas (52). In each graph, the number of mutations mapping to genes greater than 0.6 Mb in size and the total number of mutations is indicated (separated by /), followed by the statistical probability \( P \) for rejecting the null hypothesis for the bin of genes with genomic size greater than 0.6 Mb. For the 224 colorectal cancers, the analysis refers to the MSI-negative group (blue letters); the difference in frequencies between the MSI-positive and MSI-negative groups was also examined and was significant at a level of \( P < 6.9 \times 10^{-10} \). The cumulative genomic length of the genes greater than 0.6 Mb in size corresponds to 14.0% of the cumulative genomic length of the 19,075 protein coding genes used in our analysis (Supplementary Table S5).
The presence of a mutator phenotype in colon precancerous lesions is further supported by the correlation between polyp size and number of SNSs and the absence of a correlation between patient age and number of SNSs. In addition, the type of mutations targeting the oncogenes and tumor suppressor genes in the adenomas is also consistent with a mutator phenotype. Many of the APC mutations, which are the putative cancer-initiating events in our series, were smlNDEls, consistent with them being induced by exogenous mutagens. However, the mutations targeting the other tumor suppressor genes and the oncogenes in the same adenomas were mostly SNSs, suggesting a different mutagenic mechanism, which would be consistent with the establishment of a mutator phenotype.

A recent study has also tackled the question of a mutator phenotype in a human cancer, specifically in acute myelogenous leukemia (AML), and reached the conclusion that most of the passenger mutations observed in the leukemic cells had been acquired before oncogenic transformation (54). Exome sequencing of hematopoietic stem/progenitor cells (HSPC) from normal individuals was used to determine the number of SNSs present in nontransformed cells. Interestingly, the number of SNSs in the HSPCs correlated well with age. In individuals, who were 20- to 70-years old, between 1 and 11, SNSs were identified per HSPC, a number essentially identical to the number of SNSs present in the polyps of our series with 1 cancer-driver event. In the AML cells, the number of SNSs in the coding regions averaged about 13, which is only slightly higher than the number observed in the HSPCs. However, the AML cells were subjected to genome sequencing, which results in lower coverage than exome sequencing, and may miss cancer-associated SNSs. Nevertheless, it is clear that the number of SNSs in AML cells is not very different from the number of SNSs in normal HSPCs, thereby questioning the presence of a mutator phenotype in this tumor type. To reconcile the discrepancy with our findings, we note that the biology of AML differs significantly from the biology of solid tumors. The total number of SNSs observed in AML is typically 2 orders of magnitude lower than the number of SNSs present in most solid tumors, including colon carcinomas (33, 54). Furthermore, AML typically requires as few as 2 cancer-driver events to give rise to clinical symptoms (54). Thus, if AML
is characterized by a short-time window between acquisition of the cancer-driver events and diagnosis, it will be very hard to determine whether it has a mutator phenotype. In contrast, solid tumors remain in the precancerous state for years, as they require many more cancer-driver events to progress. For example, in several of the adenomas in our study, all of which were asymptomatic, we identified 5 cancer-driver events, counting SNSs, smINDELs, and LOH.

Some hints about the mechanisms leading to a mutator phenotype in precancerous lesions come from the type of SNSs and their distribution in the genome. Most SNSs observed in the colon adenomas were C to T transitions, and most of these transitions were observed in the context of CpG dinucleotides. As previously proposed, 5-methylcytosines can be spontaneously deaminated to form thymines (55). The resulting T:G base pair could then be repaired by base excision repair (BER) enzymes, most notably by MBD4 and TDG, which recognize the mismatch in the context of double-stranded DNA (56). However, DNA replication stress, which is prevalent in human precancerous lesions and cancers, is associated with long stretches of single-stranded DNA (ssDNA; refs. 5, 13–18). Thus, methylated cytosines that are deaminated in the context of ssDNA would not be repaired by BER resulting in C to T transitions (Fig. 5). The increased frequency of SNSs at very large genes is consistent with such a mechanism, as DNA replication stress targets preferentially the very large genes (48–50).

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: S.I. Nikolaev, S.E. Antonarakis, T.D. Halazonetis
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.K. Sotiriou, I.S. Pateras, S. Sougioultzis, M. Guipponi, V.G. Gorgoulis, T.D. Halazonetis
Writing, review, and/or revision of the manuscript: J. Saarela, S.E. Antonarakis, T.D. Halazonetis
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): I.S. Pateras, F. Santoni, S. Sougioultzis, S.E. Antonarakis
Study supervision: S.E. Antonarakis, T.D. Halazonetis

Acknowledgments
The authors thank Nicolas Roggli for help with figure preparation, Laurence Tropia and Anne Vannier for technical assistance, and Klas Wiman and Thomas Hellday for helpful discussions.

Grant Support
This work was supported by grants from the Swiss National Foundation and the European Commission (ERC) to S.E. Antonarakis and to T.D. Halazonetis and from the European Commission (FP7 Research Projects) to V.G. Gorgoulis.

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 October 7, 2012; accepted October 16, 2012; published online November 30, 2012.

References
Correction: A Single-Nucleotide Substitution Mutator Phenotype Revealed by Exome Sequencing of Human Colon Adenomas

In this article (Cancer Res 2012;72:6279–89), which was published in the December 1, 2012, issue of Cancer Research (1), some nucleotide data were mislabeled in the wild-type and mutant nucleotide columns of Supplementary Table S3. Specifically, for adenomas MI2 and MI3, column J [ref(wild-type)] lists the mutant nucleotides and column K [variant] lists the wild-type nucleotides at this position in the genome.

The mislabeling of the nucleotide data for adenomas MI2 and MI3 affected only the version of Supplementary Table S3 prepared for publication. In the mutation database used for the data analysis, the nucleotide data were correctly labeled. Thus, all the other data pertaining to adenomas MI2 and MI3 in Supplementary Table S3, such as the amino acid substitutions, etc., are correct, as are all the other data shown in the article. None of the conclusions of the study are affected. The authors regret this error.

A corrected version of Supplementary Table S3 is now available on the online journal.

Reference

A Single-Nucleotide Substitution Mutator Phenotype Revealed by Exome Sequencing of Human Colon Adenomas
