Supporting Information

Starr et al. 10.1073/pnas.1018012108

SI Materials and Methods

Mice. Apc^{Min} mice (C57BL/6J-Apc^{Min} /J), which harbor a T→A nonsense mutation in the *Apc* gene that results in a truncated protein product, were obtained from a breeding colony at the University of Minnesota Medical School Animal Services facility. Rosa26-LsL-*SB*11 mice (backcrossed to C57BL/6J) were a generous gift from Adam Dupuy (University of Iowa, Iowa City, IA). Villin-Cre mice [B6.D2-Tg(Vil-Cre)20Syr], strain 01XE7, were purchased from the National Cancer Institute Mouse Repository. T2/Onc mice (mixture of C57BL/6J and FVB) were described previously (1). Mice were necropsied, and both normal and tumor tissues were collected by snap-freezing in liquid nitrogen or overnight fixation in 10% buffered formalin followed by 70% ethanol. PCR primer sequences were as follows: T2/Onc forward: CGCTTCTCGCTTCTGTTCGC, T2/Onc reverse: CCACCCCCAGCATTCTAGTT; Villin-Cre forward: CAAGCCTGGCTCGACGGCC, Villin-Cre reverse: CGCGAACATCTTCAGGTTCT; Rosa26-lox-stop-lox-SB11 and Rosa26-SB11 knock-in3-primer: wild-type forward: CTGTTTTGGAGGCAGGAA, wild-type reverse: CCCCAGATGACTACCTATCCTCCC, knock-in reverse: CTAAAAGGCCTATCACAAAC. Apc^{Min} mice were genotyped as described previously (2).

Histopathology and Immunohistochemistry. Histopathological analysis of tumors and adjoining normal tissue was performed on tissues that were fixed in 10% neutral buffered formalin, routinely processed into paraffin, sectioned at a thickness of 4 μm, and stained with H&E. Multiple H&E sections were obtained from tumors from the colon and duodenum. All tissues were analyzed by an American College of Veterinary Pathologists-certified veterinary pathologist (M.G.O.) from the University of Minnesota Masonic Cancer Comparative Pathology Shared Resources facility and using the standardized nomenclature of the 2003 Consensus Report and Recommendations for pathology of mouse models of intestinal cancer (3). Immunohistochemistry was performed on 4-μm formalin-fixed, paraffin-embedded sections of small intestine which were deparaffinized and rehydrated, followed by antigen retrieval using 10 mM citrate buffer, pH 6.0, in a steamer. Staining for β-catenin was performed on a Dako Autostainer using a goat anti-human β-catenin polyclonal antibody (catalog no. sc-1496; Santa Cruz) as primary antibody (after blocking endogenous peroxidase and application of a protein block), with detection by a biotinylated donkey anti-goat antibody (Jackson ImmunoResearch Laboratories) and streptavidin-linked horseradish peroxidase (Dako) using diaminobenzidine (Dako) as the chromogen. Mayer's hematoxylin (Dako) was used as the counterstain. Small intestinal adenomas from $Apc^{Min/+}$ mice were used as a positive control tissue, and for negative control slides the primary antibody was substituted with Super Sensitive Goat Negative Serum (Biogenex).

Linker-Mediated PCR. Linkers used to sequence insertions were described previously (4). Genomic DNA was digested with NlaIII (for sequencing from the right side of T2/Onc) or BfaI (for sequencing from the left side of T2/Onc) and ligated to the linker using T4 DNA ligase. A secondary digestion was performed to destroy concatamer-generated products (XhoI for right-side cloning and BamHI for left-side cloning). These two enzymes do not cut the transposon distal to the NlaIII or BfaI sites but do cut the plasmid backbone present in the transgene concatamer, effectively destroying concatamer linker amplicons. Primary PCR was performed using primers that flank the inverted repeat/direct repeat (IR/DR) sequences and the linker. Primer sequences are listed below. Primary PCR products were diluted 1:75 and used in a secondary PCR with nested primers. Secondary PCR was performed using FusA+BC+Left2°Primer or FusA+BC+Right2°Primer and FusB+linker2°Primer primers (see below). FusA and FusB are sequences required for pyrosequencing using the 454 GSFlex machine (BC, barcode). PCR products were quantified using QuantIT picogreen assay (Invitrogen) and diluted to a concentration of 200,000 molecules/μL. All tumor samples were combined and diluted by the number of samples added, for a final concentration of 200,000 molecules/μL. Samples were sequenced using the Roche Genome Sequencer FLX using 454 pyrosequencing technology (Roche Applied Science) by the University of Minnesota Biomedical Genomics Center.

Sequences (all sequences are listed in the 5′ → 3′ direction) used were

Left linker+: GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC
Left linker−: TAGTCCCTTAAGCGGAG
Right linker+: GTAATACGACTCACTATAGGGCTCCGCTTAAGGGACCATG
Right linker−: GTCCCTTAAGCGGAGCC
Linker 1°Primer: GTAATACGACTCACTATAGGGC
Linker 2°Primer: AGGGCTCCGCTTAAGGGAC
(Note, linker primers work for both the Left linker and the Right linker)
Left 1°Primer: CTGGAATTTTCCAAGCTGTTTAAAGGCACAGTCAAC
Left 2°Primer: GGACATCTACTTTGTGCATGACACAAGTC
Right 1°Primer: GCTTGTGGAAGGCTACTCGAAATGTTTGACCC
Right 2°Primer: CCACTGGGAATGTGATGAAAGAAATAAAAGC
FusA+BC+Left2°Primer: GCCTCCCTCGCGCCATCAGAATGCCGCATTTAAGTGTATGTAAACTTC
Fusion A: GCCTCCCTCGCGCCATCAG

Example barcode (each tumor is different): AATGCCGCAT.

FusA+BC+Right2°Primer: GCCTCCCTCGCGCCATCAGAATGCCGCATTAAGGTGTATGTAAACTTC
FusB+linker2°Primer: GCCTTGCCAGCCCGCTCAGAGGGCTCCGCTTAAGGGAC
Fusion B: GCCTTGCCAGCCCGCTCAG

A full list of barcodes is available upon request.

Processing of Sequence Files. PCR amplicons were generated so that the 10-bp library-identifying barcode always appeared in the beginning of the sequence in the sense orientation. Sequence quality was outstanding even up to the first base. Using a custom Perl script, we scanned positions 1–12 of all reads for the presence of the library barcode, allowing 0 or 1 mismatch. We typically found perfect matches to a single barcode at positions 1–11, with matches at 2–12 occurring rarely. We did not find barcode sequences (0–1 mismatch) anywhere else in the read sequences. We successfully assigned 98% of all sequence reads to a library barcode. All barcodes differed by at least 2 bp, and no sequence read matched two or more barcodes in a given region. This process was carried out separately for each of the six 454 pyrosequencing regions used in these analyses.

Following barcode identification, the data for each of the six runs were merged to allow uniform handling of the entire dataset. Note that we did not attempt to assemble reads into contigs for several reasons: (*i*) The read quality was outstanding, matching the consensus with >99.9% accuracy when assembly was performed; (*ii*) the contigs tended not to tile at all, because the reads all were primed at the same location and are of similar length, conferring little advantage to using contigs; and (*iii*) assembly might introduce chimeric artifacts, particularly when two relatively closely spaced insertion sequences appeared on opposite strands.

To identify and remove IR/DR and linker sequences from each read, we applied EMBOSS Vectorstrip (5) with custom-designed modifications for pipeline application and assessed the best mismatch parameters to use. We sequentially attempted to match both construct elements (IR/DR and linker) in sense and antisense orientations with four successively less stringent parameter sets:

i) 10% mismatch allowed, long-construct elements (17–32 bp)
ii) 10% mismatch allowed, short-construct elements (<26 bp)
iii) 15% mismatch allowed, short-construct elements (<26 bp)
iv) 20% mismatch allowed, short-construct elements (<26 bp)

Note that the short-construct elements affect only the IR/DR sequence, because the linker element already is shorter than 26 bp.

As we reduced the stringency, more construct elements could be detected, but the risk of finding spurious chance matches increased. To guard against the introduction of spurious matches, we used generic scripts to assign a label to each recognized construct that was encountered from the 5′end to the 3′ end. We assigned the following labels:

Ideal: a construct with both IR/DR and linker elements in the same orientation
No-linker: a construct with an IR/DR but missing a linker
No-IR/DR: a construct missing the IR/DR
Bad: a construct with no recognizable elements or multiple IR/DRs
Unknown: anything else

For example, an ideal construct might look like 11-[+IR/DR]-42-[+linker]-4, where, in this case, the 42-bp insertion is in the sense orientation (as indicated by the + signs for IR/DR and linker). The numbers between elements in this representation indicate the number of base pairs between elements. Insertions <16 bp were considered empty or unmappable.

Ideal and nearly ideal construct configuration counts were monitored as the stringency parameters were relaxed. We expected the number of ideal configurations detected to increase and the unknown count to remain steady until we hit a parameter set that was too lax. We started with the strictest set of parameters (stringency level 0) and collected all ideal configurations. Then we took all sequences with nonideal configurations through stringency levels 1–3 to see if we could move them into the ideal category. In our final merged summary sequence table, the best status (ideal > no-linker > no-IR/DR > bad) was reported for each sequence along with the data for the most stringent run level (0–3) where it achieved this label.

Mapping Insertion Sequences. To map sequences to the mouse genome [National Center for Biotechnology Information (NCBI) Build 37], we used BLASTN (DeCypher's TeraBLASTN, Active Motif, http://timelogic.com), requiring query sequences to align within 1 bp of the start of right-IR/DR sequences or within 1 bp of the end of left-IR/DR sequences (i.e., within 1 bp of the transposon insertion site with both types of reads). Additionally, the query was required to match with at least 95% identity. Lower thresholds of 90% and 85% identity were tested but failed to yield sufficiently higher percentages of newly mappable insertions to warrant lowering the matching stringency. Because we were most interested in the IR/DR position, we were careful to ensure that the query matched within 1 bp of the IR/DR insertion site, but we did not require the 3′ end of the query to match, in case cloning artifacts had altered that end of the sequence. If secondary genome hits were found that were at least 95% as long as the first match, their count was recorded, and the insertion location was considered ambiguous. However, if all secondary hits appeared within 5,000 bp of the primary hit on the same chromosome, we considered the insertion to be uniquely mappable to that locus.

Of the 324,898 sequences analyzed by BLASTN, 173,101 (53%) could be uniquely mapped to the mouse genome. We removed redundant sequences that arose from the same tumor and mapped to the same TA dinucleotide insertion site in the genome. Of the 173,101 mapped sequences, 100,171 (67%) were redundant, leaving 72,930 nonredundant mapped insertions. Three more filtering steps were performed:

i) To avoid the bias of "local hopping," all nonredundant insertions mapping to the chromosome containing the transposon donor concatamer (Chr 1) were removed.
ii) The T2/Onc transposon contains sequence from an intron and splice acceptor of the murine En2 gene. Any insertions mapping to this sequence were removed because they might represent a transposition event back into the concatamer and not an insertion into the genomic En2 locus.
iii) Three-primer PCR was used to validate a sampling of nonredundant mapped insertions. We were successful in validating mapped insertions except in one circumstance. When a single TA dinucleotide contained multiple insertions from several tumors from multiple mice, we could not validate the insertions with three-primer PCR in some cases. We hypothesize that some of these TA dinucleotides are near sequences that cause PCR artifacts and do not represent true transposon insertions. We chose to adopt a conservative approach to avoid these artifacts, so we eliminated all insertions in TA dinucleotides containing insertions from two or more tumors from two or more different mice.

These three filtering steps removed 42,842 (59%) insertions, leaving a total of 30,088 nonredundant mapped insertions. This set was used to determine CIS.

Statistics Used to Identify CIS. To identify CIS, nonredundant insertions were assigned to clusters if the local density of insertions in a given window size exceeded that which would be expected by chance. Window sizes were determined by exact Monte Carlo simulation (see below). Based on a dataset of 30,088 insertions, the significance thresholds obtained are

Five or more insertions within 12,000 bp
Six or more insertions within 22,000 bp
Seven or more insertions within 34,000 bp
Eight or more insertions within 50,000 bp,
Nine or more insertions within 65,000 bp
Ten or more insertions within 82,000 bp
Eleven or more insertions within 105,000 bp
Twelve or more insertions within 124,000 bp
Thirteen or more insertions within 150,000 bp
Fourteen or more insertions within 175,000 bp
Fifteen or more insertions within 200,000 bp

The assumption of standard Poisson statistics that potential insertion sites are randomly distributed throughout the genome is not strictly correct, because (*i*) TA dinucleotides are naturally clustered in genomes, and (*ii*) numerous unfinished regions in

the mouse genome are "off-limits" because they are long tracts of Ns. For example, the initial telomeric region of every chromosome except Y is padded with 3 million consecutive Ns. Both these factors lead standard analytical approaches to underestimate the size and number of clusters that actually would be encountered by simply picking randomly chosen real TA sites. In other words, by ignoring the natural clustering of TA sites in the genome, the number of false-positive CIS that will be predicted is increased systematically. The magnitude of deviation gets larger as more and more insertion sites are scattered about the genome, as one would expect intuitively. Hence, we wrote a program to compute exactly the expected number of CIS of a given size in a specified window across all the chromosomes that one would encounter by chance via Monte Carlo simulation. The observed number of unambiguous mappable nonredundant insertions was used for each chromosome separately as input. For example if chromosomes 1 and 2 had 2,100 and 1,420 insertions, respectively, then we randomly distributed 2,100 insertions among the real TA dinucleotide sites on mouse chromosome 1 and another 1,420 among the TA sites of chromosome 2. Once the total count of insertions was distributed randomly among the real TA sites across the whole genome, a tally of the number of CIS of size ≥3, ≥4, …, ≥15 was recorded within windows of 10,000 bp, 20,000 bp, … 150,000 bp. This process was repeated 100 times, and the average counts over those 100 iterations were computed. Four independent simulations of 100 iterations each were performed, yielding SE bars between simulations of <1%, indicating sufficient convergence. The values obtained can be interpreted as expected values (E-values), because they indicate the expected number of CIS of a given number of insertions that would be observed within a given window size merely by chance. We chose a threshold with an E-value <1. Thus, finding 11 CIS of ≥15 insertions within 200,000 bp, when not even a single CIS was expected, is highly significant. We compared the thresholds obtained by this method with the thresholds obtained using standard Poisson statistics with the assumption of random insertion in the genome. We found our method was uniformly more stringent and yielded fewer false positives than the standard Poisson statistics. A modified Poisson model that takes into account the local density of TA sites in the genome yielded excellent agreement with the Monte Carlo calculations.

Annotation and Sequence Information Management. We created two primary annotation files: one outlining details of each unique insertion, and one describing each CIS. These files provide information on the chromosomal mapping position of each insertion or CIS, redundancy information on each insertion, and characteristics of the nearest EnsEMBL gene that flanks the insertion. EnsEMBL mappings were identified by a custom Perl script that uses the published Application Programmer Interface (6). To facilitate the management of all sequence information, an MySQL relational database was constructed to store (*i*) genotypic and phenotypic information on all mice and tumors from which the insertion sequences were derived; (*ii*) metainformation on the sequencing runs themselves; (*iii*) raw read sequences; (*iv*) construct element matching characteristics; (*v*) final processed insertion sequences; (*vi*) mapping information for each processed insert sequence to the mouse genome; (*vii*) clustering assignments of inserts into CIS; and (*viii*) annotation information on all mapped inserts and CIS. SQL queries were performed to facilitate the merging of distinct gastrointestinal tract tumor datasets and the annotation process.

Analysis of Replicate Sequencing Runs to Determine Percentage of Library Capture. To estimate the extent of undersampling of transposon insertions in our study, we analyzed the GS FLX sequencing replicates separately. Using a four-region plate, one sequencing run of the GS FLX machine can sequence four separate samples. In one of the sequencing runs we ran two aliquots of the left-side ligation-mediated (LM)-PCR pool in separate regions (1 and 4) and two aliquots of the right-side LM-PCR in separate regions (2 and 3). We analyzed the sequence reads by custom Perl script to determine the extent of overlap (Table S2). For example, region 1 returned 68,371 total reads, of which 56,435 contained a perfect match to a barcode, the transposon-specific sequence, and a genomic TA dinucleotide along with at least 16 bases of genomic DNA. When duplicate sequences were combined, 20,654 unique transposon insertion reads remained. Region 4, which was a replicate, had 18,863 unique reads. The overlap between these two regions was 10,448 reads, a little more than 50%. Because this process is similar to a mark-recapture experiment, one can use the Lincoln–Peterson method (7) to estimate the number of amplicons in the total population (see below). Based on the overlap between the replicates sampled in regions 1 and 4, we estimate that there were 37,289 unique amplicons in the original pool. As a rough approximation, our protocol sampled about 78% of the amplicons present in the original left-side LM-PCR pool. If we apply the same analysis to the right-side LM-PCR pool, where the overlap was lower (~35%), we estimate that we sampled only 58% of the amplicons in the original right-side LM-PCR pool. In either case, it is apparent that by increasing the number of sequencing runs, or perhaps by using a different sequencing platform, we could find more transposon insertions and perhaps more candidate cancer genes.

Lincoln–Peterson Method to Estimate Total Population in a Mark-and-Recapture Experiment. A mark-and-recapture experiment can be used to estimate population size in an ecological setting where the researcher marks all animals captured during a first visit. The researcher then returns and makes a second capture, noting how many of the animals in the second capture were marked in the first capture. To use the Lincoln–Peterson method (7) in our study, we assume that the duplicate pools of amplicons contain equal numbers of the same amplicons. This assumption could be invalid, because the concentration in the amplicon pool is very small (~200,000 molecules/μL). Randomly removing a small volume could result in the two volumes containing different amplicons. Another caveat to using the Lincoln–Peterson method is that our use of unique reads instead of actual reads is not directly analogous to a mark-and-recapture experiment. Nevertheless, if we assume the aliquots are similar, then the number of reads in the first region (M) is analogous to the number of animals caught and marked in the first capture of a mark-and-recapture experiment. The number of reads in the second region (C) is analogous to the number of animals caught in the second capture, and the number that overlaps (R) is analogous to the number of animals marked in the first capture that are caught in the second capture. To estimate the total population (N), the Lincoln–Peterson method states that the proportion of marked individuals in the second capture to the number in the first capture (R/M) should equal the proportion of the number of animals in the second capture to the total population (C/N). Rearrangement of this equation gives $n = (M \times C)/R$. Using the number of unique reads in regions 1 and 4 (Table S2) equates to $(20{,}654 \times 18{,}863)/10{,}448 = 37{,}289$. The total number of unique reads we captured in both sequencing regions ($20{,}654 + 18{,}863 - 10{,}448 = 29{,}069$) represents 78% of the estimated total population of 37,289. Using the same analysis of the unique reads in region 2 and 3 (Table S2) equates to an estimated total population of $(25{,}753 \times 25{,}804)/9{,}079 = 73{,}194$. In this case, we have sampled only ($25{,}753 + 25{,}804 - 9{,}079 = 42{,}478$) 58% of the total population.

***Apc* Loss of Heterozygosity Analysis.** To measure loss of heterozygosity (LOH) for the Apc^{Min} mutation, DNA was isolated from

individual polyps, and PCR was performed using primers that flank the mutation (sense primer: CGGAGTAAGCAGAGACACAA; antisense primer: GGGAGGTATGAATGGCTGAT). The PCR product was purified using Qiagen 96 MinElute vacuum purification plates according to the manufacturer's protocol and was sequenced using the sense primer as the sequencing primer. Trace peak heights at the location of the mutation were measured for each tumor. Wild-type DNA will only have a single peak, representing thymidine, whereas Apc^{Min} nontumor DNA will have two peaks of equal height representing the wild-type thymidine base and the mutant adenine base. If significant LOH has occurred, the ratio of the height of the thymidine peak to the adenine peak will be <0.5, whereas a ratio >0.8 indicates maintenance of heterozygosity. Ratios between 0.5 and 0.8 and ratios >1.2 are considered to be contaminated with nontumor tissue.

Chromosomal Copy Number Analysis. Eight studies measuring DNA copy number in CRC compared with normal tissue were analyzed (8–15). Of the 33 CIS loci, 31 were mapped to the homologous human region using the Batch Coordinate Conversion (LiftOver) utility from the University of California, Santa Cruz (UCSC) genome browser (16). The genome build appropriate to each study was analyzed. A CIS locus was determined to be lost or gained recurrently based on the individual study methodology.

To determine the chances of this overlap occurring randomly, we ran 10,000 simulations using randomly generated CIS lists compared with the recurrently lost regions in Nakao et al. (10). In this study they hybridized genomic DNA from 135 human CRC along with normal lymphocytic DNA to a BAC array (HumArray1.14, University of California San Francisco; UCSF) and used Spot/Sproc analysis software (UCSF) to determine log2 ratios. They published an Excel file with the log2 ratios for each BAC clone for each human sample. Chromosomal losses were based on a log2 ratio lower than −0.225. The genomic coordinates of the BAC clones were downloaded from the UCSF website (http://cancer.ucsf.edu/_docs/cores/array/analysis/HA1.14_clonepos_May04.20060811.txt) and were mapped to the Human genome build HG19 using the Batch Coordinate Conversion utility from the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver). After BAC clones that did not have at least 65% of samples giving a reading and positions that did not map to the HG19 build were removed, log2 ratios for 2,075 BAC clones remained. We selected the subset of these clones that had log2 ratios lower than −0.225 in at least 5% of the samples. Genomic regions recurrently lost were defined based on the genome coordinates of consecutive BAC clones that had log2ratios lower than −0.225. For example, if three adjacent clones all showed a loss, we assigned the region based on the start coordinate of the first clone and the end coordinate of the last clone. If the neighboring clones on both sides of a BAC were not decreased, the region was defined as the coordinates of the single BAC. Next, we wrote a Perl script that created 10,000 random sets of 31 CIS using the UCSC Golden Path coordinates. Each set of 31 CIS contained 31 randomly generated genomic regions that were the same length as the actual Apc^{Min} dataset CIS. The Perl script then counted the overlap between the random CIS list and the Nakao recurrently lost regions, and the process was repeated 10,000 times. In 10,000 simulations, only three random CIS lists had an overlap of 22 regions or more.

Knockdown of CIS Genes Using siRNA in SW480 Cells. SW480 cells were purchased from American Type Culture Collection. Cells were maintained in DMEM supplemented with 10% FBS, 2 mmol/L glutamine, 100 U/mL penicillin, and 100 μg/mL streptomycin and were incubated in a humidified atmosphere of 95% air and 5% CO_2 at 37 °C. Cells were switched to antibiotic-free medium before siRNA transfection. Cells at 70% confluence were transfected twice, 48 h apart, with siRNA oligonucleotides (siRNA oligos) targeting the human ortholog of the CIS candidate gene (CIS siRNA) or with a nontargeting control (control siRNA), at a final concentration of 50 nmol/L, using Lipofectamine 2000 transfection reagent. Oligos targeting CIS genes were obtained from Qiagen (Hs_ATF2_3, Hs_PDCD6IP_5 Hs_SFI1_7, and Hs_PDE4DIP_15) and from Dharmacon (CNOT1 On-TARGETplus SMARTPool). The control siRNA oligo (OnTARGETplus Nontargeting siRNA#3) was obtained from Dharmacon. Total RNA was harvested from CIS siRNA- and control siRNA-treated cells 2 or 3 d after the second siRNA transfection using the RNeasy kit (Qiagen). For each experimental sample, 1.5 μg RNA was converted to cDNA with random nonomer primers (Integrated DNA Technologies) and recombinant Omniscript Reverse Transcriptase using the Omniscript RT kit (Qiagen) according to manufacturer's instructions. CIS cDNA and 18S ribosomal cDNA (used as internal control) were amplified from total cDNA by PCR.

PCR was carried out in 10 μL using 2 μL cDNA from a 10× dilution of the 20 μL RT reaction (equivalent to 15 ng reverse-transcribed RNA) for CIS genes and a 100× dilution for 18S rRNA, 500 nM of each primer, and 5 μL LightCycler 480 SYBR Green 1 Master Mix (Roche Applied Science). PCR was performed on a LightCycler 480 System (Roche Diagnostics) in 96-well plates using the amplification protocol: one cycle of preincubation, 5 m at 95 °C; 45 cycles of amplification each consisting of denaturation at 95 °C for 5 s, annealing at 60 °C for 5 s, and elongation at 72 °C for 10 s; one cycle melting at 95 °C for 5 s, 65 °C for 1 m, heating to 97 °C; one cycle cooling at 40 °C for 30 s. Water was used as a template for negative control amplifications for each PCR run. All reactions were performed in duplicate. Standards were generated by reverse transcription of total RNA from untreated cells followed by PCR amplification to generate template DNA of the same sequence as the predicted CIS gene or 18S gene PCR product. Serial dilutions of template DNA were amplified in parallel with experimental samples and used to generate a standard curve for each gene. Data were analyzed using Roche LightCycler 480 software, and crossing points (CPs) were calculated using the absolute quantification-second derivative maximum method. The standard curve was used to determine efficiency of PCR amplification (E) for each gene. Relative mRNA levels represent the expression of the CIS gene in CIS siRNA-treated cells relative to expression in control siRNA-treated cells. CIS gene expression was normalized to 18S rRNA levels, and relative mRNA levels were calculated as described by Pfaffl (17): relative mRNA levels = $([E_{target}]^{\Delta CPtarget(control-treated)})/([E_{ref}]^{\Delta CPref(control-treated)}) \times 100$, where E is real-time PCR efficiency and CP is defined as the point at which fluorescence rises appreciably above background.

Primer sequences are as follows: activating transcription factor-2 (ATF2), 5′-TGACCGAAAGGATCATGAACTA-3′ and 5′-GCAGTCCTTTCTCAAGTTTCCA-3′; CCR4-NOT transcription complex, subunit 1 (CNOT1), 5′-CTTTCAACCCCCAATCAGACC -3′ and 5′-AGGTTTCATCTTACTCTGCTGGA-3′; programmed cell death 6-interacting protein (PDCD6IP), 5′-AGGTGTTCCCTGTCTTGGCTGC-3′ and 5′-TTCATCATAGCGAGATGCCACTGTTT-3′; phosphodiesterase 4D-interacting protein (PDE4DIP), 5′-GAGAACTCCAGGACAAGAAACAGCAT-3′ and 5′-GGATTCCTCCTGCAGAAGCTGG-3′; Sfi1 homolog, spindle assembly associated (SFI1), 5′-AGCAGCAGGAGATGAGGAACAAG-3′ and 5′-CGAACAACCACGTAGATCAACCAG-3′.

Cell Viability Assay. Viability of siRNA-treated SW480 cells was determined using a 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay (Cell Viability Kit 1; Roche Applied Sciences) in which absorbance at 595 nm is proportional to viable cell number. One day after the second siRNA transfection, cells were replated into 96-well plates at 1,250 cells per well (ATF2, PDCD6IP, and PDE4IP) or 5,000 cells per well (CNOT1) in

triplicate in regular growth medium. Cell viability was determined on days 2–6 after the second transfection. Day 6 A595 values for each treatment were normalized to day 2 values for the same treatment to correct for differences in plating. Relative cell viability represents the ratio of normalized absorbance at 595 nm of cells treated with CIS siRNA to absorbance of cells treated with control siRNA on day 6. Results shown for mRNA and cell viability are the mean ± SD of at least two experiments.

1. Collier LS, Carlson CM, Ravimohan S, Dupuy AJ, Largaespada DA (2005) Cancer gene discovery using *Sleeping Beauty* transposon-based somatic mutagenesis in the mouse. *Nature* 436:272–276.
2. Dietrich WF, et al. (1993) Genetic identification of Mom-1, a major modifier locus affecting Min-induced intestinal neoplasia in the mouse. *Cell* 75:631–639.
3. Boivin GP, et al. (2003) Pathology of mouse models of intestinal cancer: Consensus report and recommendations. *Gastroenterology* 124:762–777.
4. Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science* 300:1749–1751.
5. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
6. Stabenau A, et al. (2004) The Ensembl core software libraries. *Genome Res* 14: 929–933.
7. Seber GAF (2002) *The Estimation of Animal Abundance and Related Parameters* (Blackburn, Caldwell, NJ) Ed 2.
8. Vogelstein B, et al. (1989) Allelotype of colorectal carcinomas. *Science* 244:207–211.
9. Ried T, et al. (1996) Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer* 15:234–245.
10. Nakao K, et al. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* 25:1345–1357.
11. Habermann JK, et al. (2007) Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes Chromosomes Cancer* 46:10–26.
12. Lassmann S, et al. (2007) Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. *J Mol Med* 85:293–304.
13. Shih IM, et al. (2001) Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer Res* 61:818–822.
14. Derks S, et al. (2008) Integrated analysis of chromosomal, microsatellite and epigenetic instability in colorectal cancer identifies specific associations between promoter methylation of pivotal tumour suppressor and DNA repair genes and specific chromosomal alterations. *Carcinogenesis* 29:434–439.
15. Kurashina K, et al. (2008) Chromosome copy number analysis in screening for prognosis-related genomic regions in colorectal carcinoma. *Cancer Sci* 99:1835–1840.
16. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
17. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 29:2003–2007.

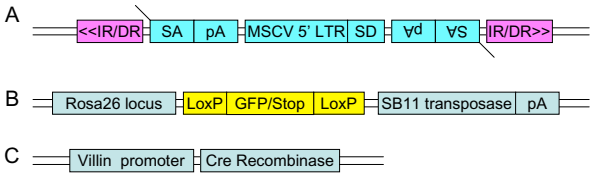

Fig. S1. Three alleles used to target *SB* mutagenesis to the intestinal tract. (*A*) T2/Onc transposon. IR/DR, inverted repeat/direct repeat sequences required for transposition; MSCV 5′ LTR, murine stem cell virus 5′ long terminal repeat; SA pA, splice acceptor with a polyA signal. (*B*) Conditional *SB* allele. Rosa26 locus, endogenous Rosa26 locus [Gt(Rosa)26Sor]; LoxP-GFP/Stop-LoxP, GFP cDNA flanked by LoxP sites; *SB*11 transposase pA, SB11 transposase cDNA with polyA signal. (*C*) Cre recombinase cDNA driven by the Villin promoter.

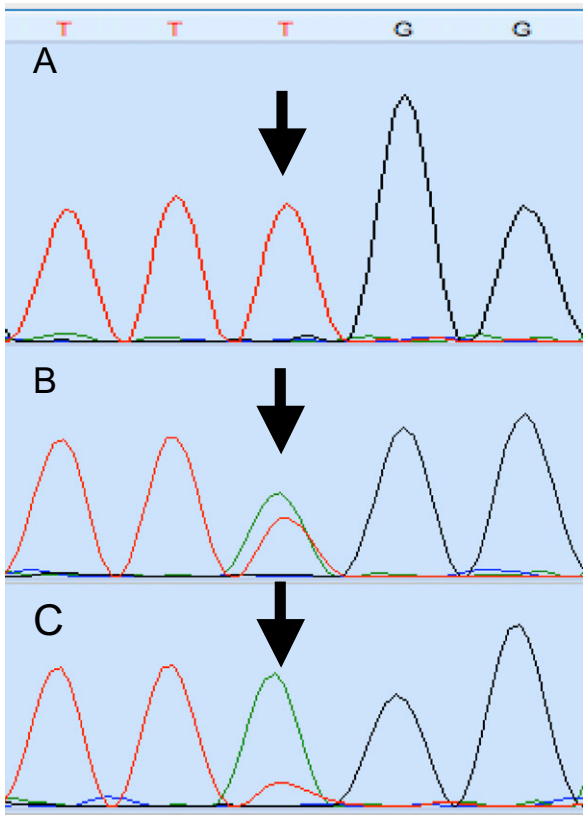

Fig. S2. PCR technique for detecting *Apc* LOH. Trace peak heights of the T→A *Min* mutation in (*A*) wild-type mice, (*B*) heterozygous Apc^{Min} mice, and (*C*) an adenoma with LOH. Arrows indicate *Min* mutation. Red peaks = T; green peaks = A.

Table S1. LOH and MOH in Apc^{Min} tumors based on the ratio of T:A trace peaks

Transposon insertion in *Apc*	Number of informative tumors*	% tumors with T:A ratio <0.5 (LOH)	% tumors with T:A ratio between 0.8–1.2 (MOH)
Yes†	32	47	53
No	15	73	27

MOH, maintenance of heterozygosity.
*To qualify as an informative tumor, the peak height ratio T:A must be <0.5 or between 0.8–1.2.
†Yes indicates the tumors had a mapped *SB* transposon insertion in the *Apc* gene.

Table S2. Sequence read overlap between duplicate regions of a single GS FLX sequencing run*

	Region 1 (left)†	Region 4 (left)†	Region 2 (right)†	Region 3 (right)†
Average read length (bases)	95	94	94	94
Total no. reads	68,371	59,228	54,405	113,982
No. reads with a perfect barcode‡	66,607	57,631	53,168	111,016
No. reads with IR/DR+TA match§	56,435	49,006	49,531	102,534
No. reads with unique 16 bases after TA¶	20,654	18,863	25,753	25,804
No. unique reads that overlap	10,448	10,448	9,079	9,079
Percent unique reads that overlap	51	55	35	35

*A four-region plate was used in the sequencing run.
†Left-side LM-PCR duplicate aliquots were run in regions 1 and 4; right-side duplicates were in regions 2 and 3.
‡The first 10 bases of the read perfectly matched one of the 96 barcodes.
§The bases immediately following the barcode perfectly matched the nested primer, the remainder of the IR/DR, and a TA dinucleotide.
¶The 16 bases immediately following the TA dinucleotide were unique compared with all other sequence reads.

Table S3. Human orthologous regions to the mouse CIS with recurrent chromosomal copy number changes based on published data

Mouse address*	Mouse gene	Mouse Entrez ID	Human address†	Human gene	Human Entrez ID	Human band	In region of loss/gain‡
chr9:3001410–3030207	AC131780.5	114673?	Deleted in humans	No homology	N/A	N/A	N/A
chrY:2781406–2897989	No Gene Y	N/A	Deleted in humans	No homology	N/A	N/A	N/A
chr3:127241424–127431287	4930422G04Rik	71643	chr4:113299994–113576517	C4orf21	55345	4q25	Yes
chr3:127697338–127730035	AC115907.7	100043382	chr4:112926160–112964108	No Gene 3	N/A	4q25	Yes
chr13:105093822–105104050	Adamts6	108154	chr5:64747713–64758225	ADAMTS6	11174	5q12.3	Yes
chr3:127454370–127633377	Ap1ar	211556	chr4:113034449–113260728	AP1AR	55435	4q25	Yes
chr18:34324389–34514767	Apc	11789	chr5:111979623–112236985	APC	324	5q22.2	Yes
chr17:80239637–80306988	Atl2	56298	chr2:38501789–38613026	ATL2	64225	2p22.2–22.1	Yes
chr18:61726208–61750047	Csnk1a1	93687	chr5:148873515–148912809	CSNK1A1	1452	5q32	Yes
chr18:73903365–73913215	Elac1	114615	chr18:48495579–48509486	ELAC1	55520	18q21.2	Yes
chr3:136918543–137092521	Emcn	59308	chr4:101323972–101541428	EMCN	51705	4q24	Yes
chr18:10578932–10773953	Esco1	77805	chr18:19121974–19393499	ESCO1	114799	18q11.2	Yes
chr3:122217801–122319064	Fnbp1l	214459	chr1:93916974–94045997	FNBP1L	54874	1p22.1	Yes
chr7:135180978–135240752	Itgam	16409	chr16:31267174–31289751	ITGAM	3684	16p11.2	Yes
chr18:74613029–74796289	Myo5b	17919	chr18:47508944–47707289	MYO5B	4645	18q21.1	Yes
chr18:26169184–26282443	No Gene 18	N/A	chr18:35353815–35494000	No Gene 18	N/A	18q12.2	Yes
chr4:131170124–131238909	No Gene-4	N/A	chr1:29768033–29870348	No Gene 4	N/A	1p35.3	Yes
chr13:55352993–55372133	Nsd1	18193	chr5:176657455–176671302	NSD1	64324	5q35.3	Yes
chr3:97593369–97718572	Pde4dip	83679	chr1:144945932–145112522	PDE4DIP	9659	1q21.1	Yes
chr11:62203495–62369763	Pigl	327942	chr17:16040912–16299937	PIGL	9487	17p11.2	Yes
chr6:113039133–113100176	Setd5/Lhfpl4	72895/269788	chr3:9450610–9516832	SETD5 LHFPL4	55209 375323	3p25.3	Yes
chr11:3004743–3179859	Sfi1	78887	chr22:31790517–32022116	SFI1	9814	22q12.2	Yes
chr18:53440600–53638364	Snx24	69226	chr5:122221887–122440150	SNX24	28966	5q23.2	Yes
chr18:52654288–52675270	Srfbp1	67222	chr5:121370968–121397902	SRFBP1	153443	5q23.1	Yes
chr18:24110107–24157877	Zfp397	69256	chr18:32819631–32871403	ZNF397	84307	18q12.2	Yes
chr9:65561432–65607064	Zfp609	214812	chr15:64859323–64946983	ZNF609	23060	15q22.31	Yes
chr2:73708689–73773260	Atf2	11909	chr2:175997880–176076969	ATF2	1386	2q31.1	No
chr8:98254541–98300340	Cnot1	234594	chr16:58568279–58622838	CNOT1	23019	16q21	No
chr16:29019897–29031672	No Gene 16	N/A	chr3:192750830–192764409	No Gene 16	N/A	3q29	No
chr9:113560446–113723693	Pdcd6ip	18571	chr3:33684948–33911522	PDCD6IP	10015	3p22.3	No
chr9:100583003–100698850	Stag1	20842	chr3:136220568–136384331	STAG1	10274	3q22.3	No
chr5:126077980–126153976	Tmem132b	208151	chr12:125776600–125970181	TMEM132B	114795	12q24.31–24.32	No
chr18:7855248–8046126	Wac	225131	chr10:28820138–28913566	WAC	51322	10p12.1	No

*Based on mouse July 2007 Assembly (mm9) UCSC genome browser.
†Based on human GRCh37 Assembly (hg19) UCSC genome browser.
‡Region was identified in one of the following studies: Nakao et al. (1); Derks et al. (2); Lassmann et al. (3); Habermann et al. (4); Ried et al. (5); Vogelstein et al. (6); Shih et al. (7).

1. Nakao K, et al. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* 25:1345–1357.
2. Derks S, et al. (2008) Integrated analysis of chromosomal, microsatellite and epigenetic instability in colorectal cancer identifies specific associations between promoter methylation of pivotal tumour suppressor and DNA repair genes and specific chromosomal alterations. *Carcinogenesis* 29:434–439.
3. Lassmann S, et al. (2007) Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. *J Mol Med* 85:293–304.
4. Habermann JK, et al. (2007) Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes Chromosomes Cancer* 46:10–26.
5. Ried T, et al. (1996) Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer* 15:234–245.
6. Vogelstein B, et al. (1989) Allelotype of colorectal carcinomas. *Science* 244:207–211.
7. Shih IM, et al. (2001) Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer Res* 61:818–822.

Table S4. Knockdown of *Apc*^{Min} CIS candidate genes affects viability of human colon cancer cells

Gene	Relative mRNA levels of cells depleted for CIS gene*	Relative cell viability of cells depleted for CIS gene†
ATF2	17.2 ± 5.6%	61.0 ± 1.5%
CNOT1	13.7 ± 7.4%	18.5 ± 16%
PDCD6IP	10.5 ± 2.6%	64.2 ± 28%
PDE4DIP	27.9 ± 10%	55.0 ± 19%
SFI1	40.4 ± 13.2%	60.0 ± 14%

*Relative mRNA levels calculated using the ΔΔCt method (1) and represent levels of CIS mRNA in CIS siRNA-treated cells relative to levels in control, nontargeting siRNA-treated cells. 18S RNA was used as a housekeeping control. Results shown are the mean ± SD of at least two experiments.

†Relative cell viability represents the ratio of absorbance at 595 nm of cells treated with the indicated CIS gene siRNA to absorbance of cells treated with nontargeting control siRNA. Results shown are the mean ± SD of at least two experiments.

1. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 29:2003–2007.

Other Supporting Information Files

Dataset S1. Mapped transposon insertions in 96 tumors

Dataset S1 (XLS)