SUPPLEMENTARY MATERIALS AND METHODS

DNA isolation and storage

DNA was extracted from paired blood and adenoma tissue samples using a DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA) according to the manufacturer's instructions. Quality control was performed and concentrations were determined using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). DNA samples were stored at −80°C until used. An Infinium CoreExome-24 BeadChip (Illumina, San Diego, CA) was employed for quality control and confirmation of sample identity.

Custom capture panel design

Through mining existing databases and a literature review, we designed a custom capture panel from 4 sources detailed below. First, using The Cancer Genome Atlas (TGCA) data and the Broad Institute Genome Data Analysis Center's Firehose pipeline,[1] MutSig v2.0 or MutSigCV v0.9[2] was applied to identify significantly mutated genes in the colon adenocarcinoma (COAD) and colorectal adenocarcinoma (COADREAD) data sets, with significant p values <0.05 and population frequency >0.01. Second, we queried driver genes that had been reported previously[3] in the COSMIC database[4] and selected genes with a population frequency >0.01 in large intestine tissues. Third, chromosome regions with recurrent copy number variations as determined using the GISTIC2 approach[5] were covered by manually selected genes to reach at least 1 gene per 10 million base pairs (Mb). Finally, we conducted a comprehensive literature search to identify genes that have been reported to be involved in colorectal cancer. Together, these 4 sources identified 767 genes, and the coordinates of these genes were obtained from the University of California, Santa Cruz Genome Browser (online Supplementary Table S1). When multiple transcripts exist, we chose the longest transcript. All exons from these transcripts were used as input for probe design using NimbleDesign (Roche, Pleasanton, CA). The final capture probes covered 99.4% of the target bases, and the capture target was approximately 3.6 Mb.

Library preparation, exome capture, and sequencing

Library construction and exome capture was performed at the Human Genome Sequencing Center at Baylor College of Medicine as previously described.[6] A complete library construction and exome capture protocol are available on the website of the Human Genome Sequencing Center (https://www.hgsc.bcm.edu/sites/default/files/documents/Protocol-Illumina_Whole_Exome_Sequencing_Library_Preparation-KAPA_Version_BCM-HGSC_RD_03-20-2014.pdf). In brief, 500 ng of genomic DNA was sheared into fragments with an average size of 200-300

base pairs (bp) using an S2 System (Covaris, Woburn, MA). The fragmented DNA underwent end repair using an NEBNext End-Repair Module (NEB, Ipswich, MA), 3'-adenylation using an NEBNext dA-Tailing Module, ligation of Illumina adaptors using NEB Quick Ligase Enzyme, and purification using SPRI AMPure XP beads (Beckman Coulter, Brea, CA) according to the manufacturers' instructions. A ligation-mediated polymerase chain reaction (PCR) was performed with processed fragments as a template using KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA) for 6 cycles. After amplification, SPRI AMPure XP beads were applied to purify the PCR products. The quality of the precapture library was then determined using a Bioanalyzer 2100 DNA 7500 Chip (Agilent, Santa Clara, CA).

Three micrograms of the precapture library was mixed with hybridization buffer, Cot-1 DNA (Invitrogen, Waltham, MA), and hybridization-enhancing oligos (Sigma-Aldrich, St. Louis, MO). After the mixture had been denatured for 10 minutes at 95°C, SeqCap EZ HGSC VCRome capture probes (Roche, Pleasanton, CA) were added, targeting about 37 Mb covering 23 585 genes and 189 028 exons or the aforementioned custom capture probes. The samples were incubated at 47°C for 64 to 72 h. Streptavidin Dynabeads (Invitrogen, Waltham, MA) were preheated at 47°C for 5 minutes and transferred to the hybridization reactions. After 45 min, the beads were washed, and the bound DNA was eluted. The postcapture libraries were amplified with KAPA HiFi HotStart ReadyMix for 10 to 12 cycles. The PCR products were cleaned with SPRI AMPure XP and eluted in nuclease-free water. Every 4 whole-exome or 30 custom capture libraries were pooled and sequenced in a single lane of an Illumina HiSeq2000 with 2 × 100-bp paired-end reads.

Pipeline for mapping, somatic mutation calling, and annotation

To compare and contrast the somatic mutation profiles of premalignant (adenoma) and colorectal cancer (CRC) samples, we downloaded TCGA DNA sequencing data for primary CRC tissues (N = 460) and corresponding blood samples from the Cancer Genomics Hub[7] (dbGaP Study Accession: phs000178.v9.p8). The downloaded aligned BAM files were converted to FASTQ files using the SamToFastq functionality of Picard Tools V1.118 (http://broadinstitute.github.io/picard). The FASTQ files generated from the adenoma and CRC patient samples were processed in an in-house analysis pipeline developed for the sequencing data. FASTQ files were mapped to the human reference genome HG19 using Burrows-Wheeler Aligner V0.7.10[8] to generate SAM files. Multiple scripts in Picard Tools were employed to fix mate pairs and to compress, index, and sort SAM files to BAM files. After that, local realignment around known indels identified in dbSNPv137 and the 1000Genomes Project was performed using Indel Realigner (GATK v.3.3-0).[9] Duplicates in the realigned BAM file were marked using Picard Tools, and the

base quality was recalibrated with GATK Base Recalibrator to generate analysis-ready reads. Somatic variant calls were carried out using MuTect V1.1.7[10] and VarScan V2.3.7,[11] comparing sequencing data from the paired adenoma and blood samples. MuTect was run with default parameters, and VarScan was run with 3x minimum coverage, minimum mean base quality of 15, minimum variant allele frequency of 0.05, and somatic p value < 0.05. We pooled the variant calls from both callers and applied a universal filter excluding variants with fewer than 2 copies of reads supporting the alternative base on single nucleotide variants (SNV). Indels were called with Varscan with the same parameters used for SNV but were filtered by (1) total tumor reads > 15; (2) total normal reads > 6; (3) total number of reads supporting a call > 4; (4) variant allele frequency (VAF) in tumor >5%; and (5) VAF in normal tissue <1%. Output from variant callers was converted to VCF format and annotated using the variant annotation tool in the VAAST2 package.[12]

Mutation preprocessing, subsetting, prioritization, and testing for significance

Annotated mutations were converted to Mutation Annotation Format (MAF) v2.4.1 according to National Cancer Institute specifications (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification). Mutations from TCGA CRC patients marked as "do not use" in the TCGA annotation manager were removed. Hypermutaters were identified by combining MutL homolog 1 (*MLH1*) expression and microsatellite stability status. Z scores for *MLH1* expression were measured using RNAseqV2, and microsatellite stability information was downloaded from cBioPortal[13] via the cgdsr package in R[14]. *MLH1* expression Z scores larger than 1.96, corresponding to p values < 0.05, were defined as *MLH1* dysregulation. Median raw expression was employed to further separate dysregulation into upregulation and downregulation. Both high microsatellite instability and low microsatellite instability were defined as microsatellite instability. When a patient had more than 1 mutation of the same gene, the mutation with the most severe consequence was chosen to represent the underlying mutational status of the gene. The order of mutation consequences from the most to the least severe was as follows: nonsense, splice site, missense, nonstop, silent, untranslated region, intronic, and noncoding RNA or unannotated genes. Nonsilent mutations in our analysis included nonsense, splice site, missense, and nonstop mutations. The frequency of the gene mutations was calculated from the MAF files, and previously defined false-positive genes[15] were removed. Hypermutaters were identified by mutL homolog 1 (*MLH1*) expression levels and microsatellite stability statuses obtained from the TCGA data portal. In order to identify potential driver genes, we used SomInaClust[16] under default parameters and COSMIC V71 as the reference file.

Identification of mutation signatures for adenoma and CRC

The Student *t* test was used to compare mutation rates in adenoma and colorectal cancer samples. The Fisher exact test was employed to identify differences in mutation frequency according to pathological and clinical features. Classification and regression tree analysis[17] was performed to discover differently mutated genes between sessile serrated adenoma and tubular villous or conventional adenoma samples. Supervised learning via random forest and permutation tests for variable importance was performed using the randomforest[18] and rfPermute[19] packages on the pooled dataset containing both adenoma and TCGA CRC datasets. Permutation of the random forest class labels was performed for 1000 iterations to provide a better estimate of variable importance than classic random forest. A reduced model was constructed using important variables identified in the random forest test with permuted p values < 0.05. For the pooled analysis of colorectal cancer and adenoma data, we further filtered the mutations using VCRome and custom panel probes to ensure similar coverage of variants.

References

1 Broad Institute TCGA Genome Data Analysis Center. Analysis Overview for Colon Adenocarcinoma (Primary solid tumor cohort) - 23 September 2013. Published Online First: 2013. doi:10.7908/C1N014TR

2 Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**:214–8. doi:10.1038/nature12213

3 Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. *Science* 2013;**339**:1546–58. doi:10.1126/science.1235122

4 Forbes SA, Bindal N, Bamford S, *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;**39**:D945-950. doi:10.1093/nar/gkq929

5 Mermel CH, Schumacher SE, Hill B, *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:R41. doi:10.1186/gb-2011-12-4-r41

6 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7. doi:10.1038/nature11252

7 Wilks C, Cline MS, Weiler E, *et al.* The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database J Biol Databases Curation* 2014;**2014**. doi:10.1093/database/bau093

8 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* 2009;**25**:1754–60. doi:10.1093/bioinformatics/btp324

9 Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* 2013;**43**:11.10.1-33. doi:10.1002/0471250953.bi1110s43

10 Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9. doi:10.1038/nbt.2514

11 Koboldt DC, Zhang Q, Larson DE, *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**:568–76. doi:10.1101/gr.129684.111

12 Hu H, Huff CD, Moore B, *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 2013;**37**:622–34. doi:10.1002/gepi.21743

13 Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. doi:10.1126/scisignal.2004088

14 Jacobsen A. *cgdsr: R-Based API for Accessing the MSKCC Cancer Genomics Data Server (CGDS)*. 2015. https://cran.r-project.org/web/packages/cgdsr/index.html (accessed 22 Apr2016).

15 Fuentes Fajardo KV, Adams D, NISC Comparative Sequencing Program, *et al.* Detecting false-positive signals in exome sequencing. *Hum Mutat* 2012;**33**:609–13. doi:10.1002/humu.22033

16 Van den Eynden J, Fierro AC, Verbeke LPC, *et al.* SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinformatics* 2015;**16**:125. doi:10.1186/s12859-015-0555-7

17 Therneau T, Atkinson B, port) BR (author of initial R. *rpart: Recursive Partitioning and Regression Trees*. 2015. https://cran.r-project.org/web/packages/rpart/index.html (accessed 22 Apr2016).

18 Cutler F original by LB and A, Wiener R port by AL and M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. 2015. https://cran.r-project.org/web/packages/randomForest/index.html (accessed 22 Apr2016).

19 Archer E. *rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics*. 2016. https://cran.r-project.org/web/packages/rfPermute/index.html (accessed 22 Apr2016).

Supplementary Table S1. Genes in the custom capture panel for targeted sequencing. Listed in separate file due to large table size.

Supplementary Table S2. Characteristics of adenoma patients

	CNAD (n=135)	SSA (n=14)	p value*
Age			
mean	59.3	58.21	0.703
Gender			
Female	58	7	
Male	77	7	0.778
Race			
White	112	13	
Black	15	1	
Other	8	0	1
Hispanic or Latino			
Y	10	1	
N	125	13	1
Smoking			
Never	66	6	
Former	50	6	
Current	19	2	0.927
Larger than 10mm			
Y	21	3	
N	113	11	0.701
High grade dysplasia			
Y	6	0	
N	129	14	1
Advanced			
Y	30		
N	104		

***:** Student's t test for continuous variables and Fisher's exact test for categorical variables.
CNAD: conventional adenoma. SSA: Sessile serrated adenoma.

Supplementary Table S3 Characteristics of patients with colorectal cancer from TCGA

	Hypermutated (n=66)	Non-hypermutated (n=312)	p value
Age			
Mean	65.77	64.92	0.653
Gender			
Female	33	146	
Male	33	166	0.685
Race			
White	36	181	
Black	7	24	
Other	5	6	0.049
Hispanic or Latino			
Y	0	2	
N	46	202	1
Pathology T stage			
T1	3	8	
T2	13	60	
T3	43	207	
T4	7	32	
Tis	0	1	0.834
Pathology N stage			
N0	55	172	
N1	7	85	
N2	4	54	
NX	0	1	<0.001
Pathology M stage			
M0	54	232	
M1	3	44	
MX	6	34	0.093
Stage			
I	15	58	
II	39	105	
III	8	95	
IV	3	46	<0.001

***:** Student's t test for continuous variables and Fisher's exact test for categorical variables

Supplementary Table S4. Driver genes with different prevalence among CNAD and CRC. Low mutation frequency CRC (LMC): CRCs with mutation frequency equal to or less than that of CNAD**. Normal mutation frequency CRC (NMC):** CRCs with mutation frequency higher than that of CNAD but not hypermutaters**.**

Gene	Non-silent mutation	CNAD (No. of samples)	LMC (No. of samples)	NMC (No. of samples)	CNAD v.s. LMC p value*	CNAD v.s. NMC p value*	LMC v.s. NMC p value*
TP53	N	34	53	54			
	Y	1	43	162	1.59×10^{-6}	5.71×10^{-17}	3.64×10^{-7}
NHEDC1	N	30	58	91			
	Y	5	38	125	6.33×10^{-3}	1.03×10^{-6}	3.22×10^{-3}
PIK3CA	N	35	75	156			
	Y	0	21	60	9.67×10^{-4}	5.56×10^{-5}	0.328
KRAS	N	32	62	120			
	Y	3	34	96	2.02×10^{-3}	2.41×10^{-5}	0.171
APC	N	19	34	24			
	Y	16	62	192	0.070	3.68×10^{-8}	1.15×10^{-6}
FBXW10	N	33	80	144			
	Y	2	16	72	0.152	4.93×10^{-4}	2.61×10^{-3}
CDRT15	N	35	89	180			
	Y	0	7	36	0.189	3.88×10^{-3}	0.032
GOLGA8B	N	23	68	102			
	Y	12	28	114	0.669	0.047	1.25×10^{-4}
FAM153C	N	22	64	94			
	Y	13	32	122	0.683	0.044	2.15×10^{-4}
CNTNAP3B	N	19	55	68			
	Y	16	41	148	0.843	0.012	2.95×10^{-5}
NRAS	N	35	88	190			
	Y	0	8	26	0.108	0.032	0.432
CTAGE9	N	34	83	174			
	Y	1	13	42	0.111	0.014	0.260
SMAD4	N	34	84	182			

	Y	1	12	34	0.183	0.038	0.494
AMER1	N	29	89	180			
	Y	6	7	36	0.108	1	0.032
TCP10	N	23	72	119			
	Y	12	24	97	0.376	0.273	$1.01x10^{-03}$
C16orf3	N	34	95	200			
	Y	1	1	16	0.464	0.481	0.027
GOLGA8A	N	23	65	105			
	Y	12	31	111	0.836	0.070	$2.04x10^{-3}$
MUC7	N	35	96	205			
	Y	0	0	11	1	0.371	0.021
RFPL3	N	34	93	188			
	Y	1	3	28	1	0.093	$6.86x10^{-3}$

***: Fisher exact test**

Supplementary Table S5. Trend test results for significant genes in random forest models

Gene	Non-silent mutation	Non-advanced CNAD (No. of samples)	Advanced CNAD (No. of samples)	non-hypermutater CRC (No. of samples)	Trend test p value*	Trend test FDR*
TP53	N	102	28	120		
	Y	2	2	192	9.80×10^{-29}	6.38×10^{-26}
KRAS	N	99	22	184		
	Y	5	8	128	5.47×10^{-12}	1.78×10^{-9}
APC	N	72	8	102		
	Y	32	22	210	4.06×10^{-10}	8.80×10^{-8}
PIK3CA	N	104	30	239		
	Y	0	0	73	3.74×10^{-9}	6.09×10^{-7}
SMAD4	N	104	30	272		
	Y	0	0	40	2.88×10^{-5}	3.75×10^{-3}
FBXW7	N	104	26	278		
	Y	0	4	34	1.10×10^{-3}	0.109
CTNNB1	N	91	27	300		
	Y	13	3	12	1.17×10^{-3}	0.109
SYNE1	N	97	27	251		
	Y	7	3	61	1.39×10^{-3}	0.113
CDC27	N	81	23	194		
	Y	23	7	118	2.01×10^{-3}	0.145
CSMD1	N	101	27	270		
	Y	3	3	42	2.68×10^{-3}	0.175
NRAS	N	104	30	292		
	Y	0	0	20	3.89×10^{-3}	0.230
RYR3	N	103	29	287		
	Y	1	1	25	7.40×10^{-3}	0.321
NALCN	N	104	30	295		
	Y	0	0	17	7.99×10^{-3}	0.325
LRP1B	N	98	28	265		

	Y	6	2	47	$8.95x10^{-3}$	0.343
FAT4	N	99	28	270		
	Y	5	2	42	0.011	0.387
ATM	N	101	30	283		
	Y	3	0	29	0.016	0.446
TMPRSS13	N	92	27	246		
	Y	12	3	66	0.019	0.481
SOX9	N	101	29	283		
	Y	3	1	29	0.023	0.522
CSMD3	N	101	30	286		
	Y	3	0	26	0.031	0.634
MED12	N	103	30	297		
	Y	1	0	15	0.049	0.666

***:** Chi-square trend test

Supplementary Table S6. Summary of previous studies exploring mutations in adenoma

	Author and Date			
	Nikolaev et al., 2012	Zhou et al., 2013	Vaqué et al., 2015	Chen et al., 2016
Sequencing and sample type	WES in 1 HP, 16 CNADs, 1 SSA, 4 CRCs	Discovery WES in normal tissue, CNAD, and CRC from 1 patient. Validation of 54 SNVs by TS in 215 CRCs and 73 pairs of adenomasa and CRCs	WES in 1 HP, 8 CNADs, and 4 CRCs from 4 CRC patients	WGS in 2 CNADs and 2 SSAs. Whole-transcriptome sequencing in 7 adenomas [b]
Capture probe	SureSelect Human Exon v3 (Agilent Technologies) or SeqCap EZ Human Exome Library SR v1.2 (Roche-Nimblegen)	NimbleGen 2.1 M Human Exome Array	SureSelect Human All Exon	N/A
Sequencing technology	Illumina HiSeq2000 and GAIIx. Paired-end 105 nt .	Illumina GAII	Illumina HiSeq2000. Paired-end 75 nt	Not specified
Depth	Polyp depth 155x. Normal depth 146x	CRC depth 45.12x. CNAD depth 46.44x. Normal depth 46.69x	99x	Not specified
Bioinformatic software				
Mapping	BWA	BWA	GEM/BFAST	BWA
SNV calling	modified SAMtools score	SAMtools	SAMtools	In-house procedure
INDEL calling	Pindel	N/A	N/A	not specified
CNAD Driver				
APC	CNAD	Adenoma [a]	CNAD	CNAD, SSA
KRTAP4-5				CNAD
CTNNB1	CNAD			
GOLGA8B				
TMPRSS13				
KRAS	CNAD		CNAD	

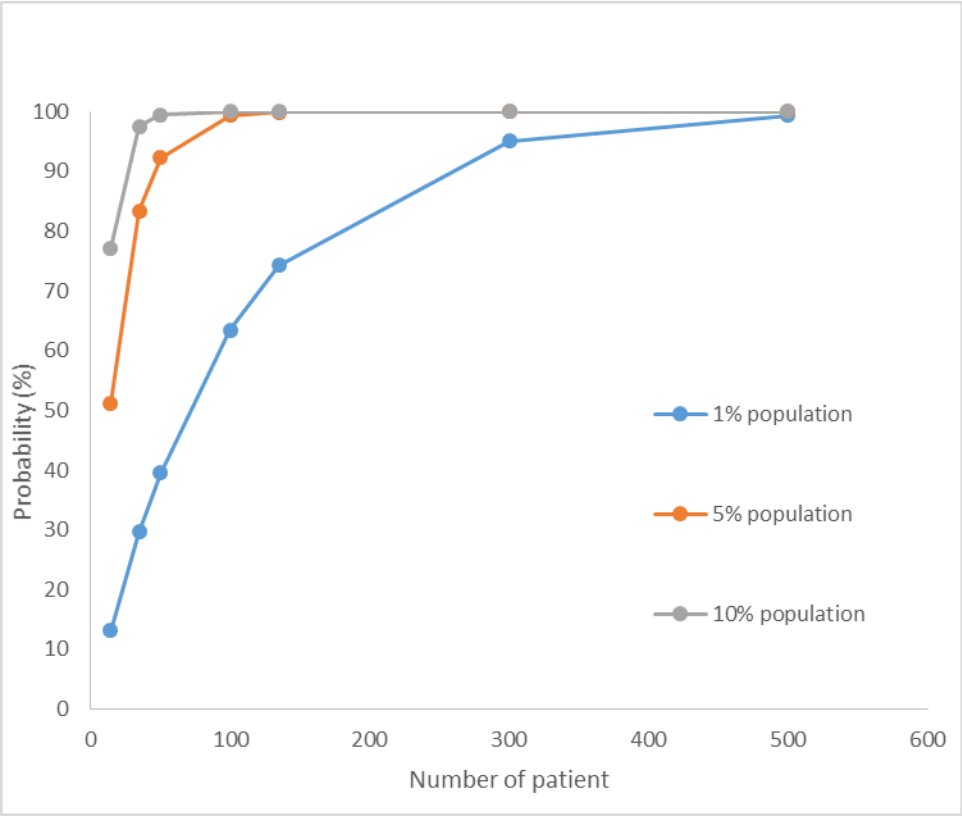
SSA driver	
BRAF	SSA

Abbreviations: WES, whole-exome sequencing; CNAD; conventional adenoma; SSA, sessile serrated adenoma; CRC, colorectal cancer; SNV, single-nucleotide variant; TS, targeted sequencing; HP, hyperplastic polyp ; WGS, whole-genome sequencing; N/A, not applicable

[a] The authors did not specify CNAD or SSA, so we use "adenoma."

[b] The comparison was made using whole-genome sequencing rather than whole-transcriptome sequencing results.

Supplementary Figure S1. Probability of detecting mutations with different prevalence. Our WES of CNADs (N = 35) was expected to detect mutations in 1%, 5%, and 10% of all CNADs with probabilities of 30%, 83%, and 98%, respectively. For WES of sessile serrated adenomas (SSAs; N = 14), the probability of detecting mutations in 1%, 5%, and 10% of these tumors was 13%, 51%, and 77%, respectively. TS provided a 63%, 99%, and 100% chance of detecting mutations present in 1%, 5%, and 10% of CNADs (N = 100), respectively.



Supplementary Figure S2. Driver mutations in colorectal cancer with mutation frequency similar to those in conventional adenoma. OG, oncogene; TSG, tumor suppressor gene.

Gene	OG mutations	TSG mutations	Driver gene p value	Driver gene q value
APC	2	89	4.45E-108	5.95E-104
KRAS	35	0	3.43E-55	4.58E-51
PIK3CA	22	0	3.94E-33	2.63E-29
TP53	30	14	6.47E-28	9.65E-21
TMPRSS13	25	0	2.16E-23	9.63E-20
NRAS	7	0	4.21E-14	1.13E-10
COA7	3	0	2.97E-09	6.61E-06
SOX9	8	10	7.11E-09	4.75E-05
KRTAP4-3	9	0	3.15E-07	0.000601
FBXW10	10	0	1.20E-06	0.002013
FBXW7	7	4	2.28E-09	0.002082
RIMBP3C	8	0	1.65E-06	0.002208
KRTAP4-5	6	0	8.35E-06	0.01015
TBC1D26	8	0	1.06E-05	0.011773
AMER1	1	5	6.50E-06	0.021723