Supporting Information

Sheffer et al. 10.1073/pnas.0902232106

SI Methods

The data we work on was acquisitioned within the framework of a Program Project Grant funded by the NCI, aimed at discovery of reliable molecular markers or targets for the detection, diagnosis, and treatment of colon cancer.

Expression Data. The expression data are composed of 390 Affymetrix GeneChip Human Genome U133A arrays. Seventeen repeats were removed from the analysis, and 2 outliers with high rates of absent values (>14,500 per sample), 9 normal colon samples whose expression profiles were markedly different from the other normals, 2 mislabeled samples, 1 microadenoma, 2 high grade adenomas, 30 noncolon normal samples and cell-lines, and 28 metastasis samples that exhibited high levels of stromal contamination (1). After all these filtering steps 299 samples were left: 180 primary carcinomas, 46 polyps, 43 normal colon epithelia, 21 liver metastases, and 9 lung metastases (see Table S3). The data were subjected to the following preprocessing steps: The Affymetrix MAS5 algorithm was applied on the .cel files (we tried also RMA and had reasons to work with MAS 5.0, see Fig. S3). Exact location for each gene was determined using UCSC hg18 known/RefSeq gene tables (2–4). The Affymetrix alignments of each probe set were compared with its chromosomal location. Probe sets (1,745) that did not have gene symbol/chromosomal location/alignments (including Affymetrix markers) or showed disagreement between their chromosomal location and alignments were removed (Affymetrix chromosomal locations were used for 432 probe sets whose genes did not appear in the University of California Santa Cruz (UCSC) known/RefSeq gene tables). Data were thresholded to T = 10 (probe sets with expression level less than T were assigned the value T). Probe sets (3,538) that had either no present calls or had expression value T in all samples, were removed. Probe sets (11,159) that represented unique genes were kept. When there were multiple probe sets per gene, the probe set that did not have _s, _x in its identifier or had the highest expression levels was chosen. Data were subjected to log2 transformation.

SNP Data. The SNP data are composed of 154 SNP-based Affymetrix 50 k XbaI GeneChip Mapping Array (5). Seven mislabeled tumor samples were removed from the analysis, and 8 outlier samples (5 normal colon, 1 normal liver, and 2 primary carcinomas) and 9 cell lines, leaving 130 samples: 62 primary carcinomas, 43 normal colon epithelia, 4 normal liver, 3 normal lung, 8 liver metastases, and 10 lung metastases (see Table S3). The preprocessing steps taken for the SNP data were as follows: The data were normalized using an algorithm developed by L. Hertzberg and O. Zuk (unpublished data), providing copy number ratio values of each allele separately, for 57,768 SNPs, using the 21 normal colon samples of female subjects from our data set as a reference set. Both alleles were summed for each SNP, data were thresholded to 1 to avoid negative or small numbers (6) and log2 transformation was performed. *Log-copy number ratios* $CR_{n,s}$ were calculated by subtracting from the log-transformed copy number of SNP n in sample s the log-transformed copy number of SNP n in the matching normal sample. If there was no matching normal, the median log-transformed copy numbers of SNP n in all normal samples were used. For chromosome X, all normal samples of the same gender were used. Smoothing was applied using the segmentation method gain and loss analysis of DNA (GLAD) (7). The $CR_{n,s}$ values were the input to GLAD and the output was a set of segments per sample, where each segment was assigned a copy number, which was applied to all SNPs in that segment of the specific sample, producing new $CR_{n,s}$ values. Since GLAD tends to misidentify outliers as separate segments (6, 8), segments with fewer than 8 markers were joined to the neighboring segment with the closest copy number, assigning the new segment the weighted mean copy number. This step was performed recursively until there were no more such small segments.

Identification of CINons. Our analysis was based on genomic identification of significant targets in cancer (GISTIC) a statistical algorithm developed by Beroukhim et al. (10). The input to GISTIC were data from 55 aneuploid tumor samples (marked in red in Fig. 1, see text). The thresholds $\theta^{amp} = 0.0974$, $\theta^{del} = -0.1121$ were used, corresponding to 0.1% quantiles of the distribution of all $CR_{n,s}$ from the normal samples. G_i scores (for each SNP) were calculated for amplifications and deletions separately, and each score was assigned a FDR q-value (9) that assesses its statistical significance, taking into consideration its amplitude and frequency, over the null distribution where the $CR_{n,s}$ in each sample are randomly permuted. The q-values were generated as follows: the SNPs were sorted according to their p-values (p_i), and the q_i-values of the procedure were corrected, in descending order, according to $q_i = \min(\min(q_i, q_{i+1}), 1)$. When the contiguous set of deleted/amplified SNPs extended over the centromere, it was divided into 2 CINons, one on each chromosomal arm. This configuration identified all broad regions and a few focal regions, and is referred to as *configuration 1*. Driver mutations are presumed to be located in the peaks [of the $-\log(q\text{-value})$] measured across the CINons, i.e., at the SNPs having the lowest q-value). Since some of the broad regions exhibit fairly uniformly low q-values and no clear prominent peak, a different configuration, referred to as *configuration 2*, was used to identify the peak regions. In this configuration, SNPs located on GLAD segments that consist of at least 75% of the SNPs in a chromosomal arm of a specific sample, were assigned copy number zero which means that these SNPs, in this sample, are removed from the subsequent analysis. These modified $CR_{n,s}$ values were then used as an input to the second run of GISTIC, keeping the same thresholds as in configuration 1. Focal/peak regions were looked for, by taking the SNPs with the lowest q-values in each of the CINons found by this second run. In this way, it was possible to identify some new focal regions and some peak regions that belonged to the broad CINons identified earlier. The peak region in 7q31.1 was manually added to the list of peak CINons as it appeared significant by eye but had the second lowest q-values. A sample s was considered to have an aberrant CINon if $median(CR_{i,s}) < \theta^{del}$ or $median(CR_{i,s}) > \theta^{amp}$, for all SNPs i located within this CINon (shown in Table 1). Leave-one-out statistics was used originally in GISTIC to determine robust borders of a CINon; this part was not performed in our analysis since some of the peak CINons are based on only 1 or 2 samples, and by removing these samples we would have missed the peak. The same analysis was performed with GISTIC over all 80 tumor samples (by adding the 25 near-diploid samples), producing almost identical results.

Correlation Between Expression and Copy Number. Denote by K the group of all colon tissue samples (including normal colon, primary carcinoma, and metastasis to lung and liver) that had both measures of expression and copy numbers (79 samples, Table S3). For each probe set n, the Pearson correlation

coefficient was calculated between the $ER_{n,s}$ values of these 79 samples, and the corresponding values $MCR_{n,s}$ that represent the copy number ratios measured in sample s from SNPs located near the probe set n. The value of each of these $MCR_{n,s}$ was determined from the median of the $CR_{i,s}$ ratios of all of the SNPs i located within a window near probe set n. The window used extends from 1,000 bps before the transcription start position to 1,000 bps after the transcription end position of the gene represented by the probe set n. If no SNPs were found in this window, the mean of the first SNP before the transcription start position and the first SNP after transcription end position was taken. We calculated the correlation only for those probe sets for which both expression and copy number shifted (versus normals) in the same direction, i.e., $mean(ER_{n,s})*mean(MCR_{n,s}) > 0$, where we average over the samples in K. Each correlation was assigned a P-value for testing the hypothesis of no correlation against the alternative, of a nonzero correlation. FDR of 25% was then used to filter the most significantly correlated probe sets.

CINon Expression Table. To asses whether a sample harbors amplification or deletion, we compared the $CE_{i,s}$ and the $CC_{i,s}$ (see *Methods*) to the upper and lower 0.1% quantiles of these values in the normal samples.

Annotation. Annotations of genes and noncoding small RNA were taken from UCSC tables, hg18 (2–4, 10, 11). Genes were considered to belong to a CINone if located in the interval between the first gene before and the first gene after the CINon's boundaries.

Outcome. Outcome labels were assigned according to the following rules: recurrence after more than 60 months was considered a good outcome, recurrence after less than or equal to 60 months was considered a poor outcome. In case there was no recurrence or recurrence information was not available, the follow-up status of the patient was tested; if the patient had died of the disease it was considered a poor outcome, otherwise, if the patient had a follow-up interval larger than 60 it was considered a good outcome. Follow-up interval of less than or equal to 60 was considered an unknown outcome. For the Kaplan-Meier survival analysis, only primary tumors were taken.

Pathway Analysis. The list of probe sets that showed significant correlation between their expression levels and their copy numbers, that were also located within the broad CINons found earlier, including the focal CINon 1p, were analyzed using DAVID (12, 13) for enrichment of Biocarta and KEGG pathways. The background was chosen as Affymetrix HG-U133A. Pathways that passed 25% FDR according to DAVID were selected.

Mutation Status. Mutation status of *p53*, *APC*, *kRAS* were obtained and also methylation status for *APC* was collected. *APC* was considered as mutated if it was either mutated or methylated.

MIN Status. Microsatellite instability (MSI) status of a tumor was determined by the National Cancer Institute (NCI) according to a set of 5 microsatellite markers demonstrating instability (14). When 2 or more markers are positive the tumor is considered as MSI-High (MIN).

***T*-Test for Genes That Are Related to the Putative TSG and Oncogenes.** Z-test was used to test the hypothesis that a value of gene i in sample j belongs to the distribution of the corresponding values of this gene in the normal samples. This was done to all of the genes in the expression data. FDR of 5% was used on the whole set of calculated p-values, resulting in a matrix of '1's, '0's, '−1's where '1′ is placed for all upregulated values that passed FDR, '−1′ for all downregulated values that passed FDR and '0′ for the rest. This matrix represents the genes that were differentially expressed in each cancer sample versus normal tissues.

For each putative tumor suppressor or oncogenes, a t test was performed comparing the samples with '1′ against samples with '0′ (for *CCDC68* the samples with '−1′ were compared against samples with '0′). FDR of 5% was used to select the genes that best separate the 2 groups of samples. Pathway enrichment analysis was performed using DAVID as described above.

1. Tsafrir D, et al. (2005) Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* 21:2301–2308.
2. Karolchik D, et al. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31:51–54.
3. Hsu F, et al. (2006) The UCSC Known Genes. *Bioinformatics* 22:1036–1046.
4. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33:D501–504.
5. Bacolod MD, et al. (2008) The signatures of autozygosity among patients with colorectal cancer. *Cancer Res* 68:2610–2621.
6. Beroukhim R, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* 104:20007–20012.
7. Hupe P, et al. (2004) Analysis of array CGH data: From signal ratio to gain and loss of DNA regions. *Bioinformatics* 20:3413–3422.
8. Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21:3763–3770.
9. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate-a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol* 57:289–300.
10. Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res* 32:D109–111.
11. Weber MJ (2005) New human and mouse microRNA genes found by homology search. *FEBS J* 272:59–73.
12. Dennis G, Jr., et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4:P3.
13. Hosack DA, et al. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol* 4:R70.
14. Boland CR, et al. (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* 58:5248–5257.

SI References for CGH, SNP studies

1. Andersen CL, et al. (2007) Frequent occurrence of uniparental disomy in colorectal cancer. *Carcinogenesis* 28:38–48.
2. Bartos JD, et al. (2007) aCGH local copy number aberrations associated with overall copy number genomic instability in colorectal cancer: Coordinate involvement of the regions including BCR and ABL. *Mutat Res* 615:1–11.
3. Camps J, et al. (2006) Genome-wide differences between microsatellite stable and unstable colorectal tumors. *Carcinogenesis* 27:419–428.
4. Camps J, et al. (2008) Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. *Cancer Res* 68:1284–1295.
5. Cardoso J, Boer J, Morreau H, Fodde R (2007) Expression and genomic profiling of colorectal cancer. *Biochim Biophys Acta* 1775:103–137.
6. Diep CB, et al. (2004) Genome characteristics of primary carcinomas, local recurrences, carcinomatoses, and liver metastases from colorectal cancer patients. *Mol Cancer* 3:6.
7. Douglas EJ, et al. (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* 64:4817–4825.
8. Grade M, et al. (2006) Aneuploidy-dependent massive deregulation of the cellular transcriptome and apparent divergence of the Wnt/beta-catenin signaling pathway in human rectal carcinomas. *Cancer Res* 66:267–282.
9. Habermann JK, et al. (2007) Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes Chromosomes Cancer* 46:10–26.
10. Jones AM, et al. (2007) Analysis of copy number changes suggests chromosomal instability in a minority of large colorectal adenomas. *J Pathol* 213:249–256.
11. Kim MY, et al. (2006) Recurrent genomic alterations with impact on survival in colorectal cancer identified by genome-wide array comparative genomic hybridization. *Gastroenterology* 131:1913–1924.
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* 85:289–300.
13. Martin ES, et al. (2007) Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. *Cancer Res* 67:10736–10743.
14. Tanami H, et al. (2005) Involvement of cyclin D3 in liver metastasis of colorectal cancer, revealed by genome-wide copy-number analysis. *Lab Invest* 85:1118–1129.

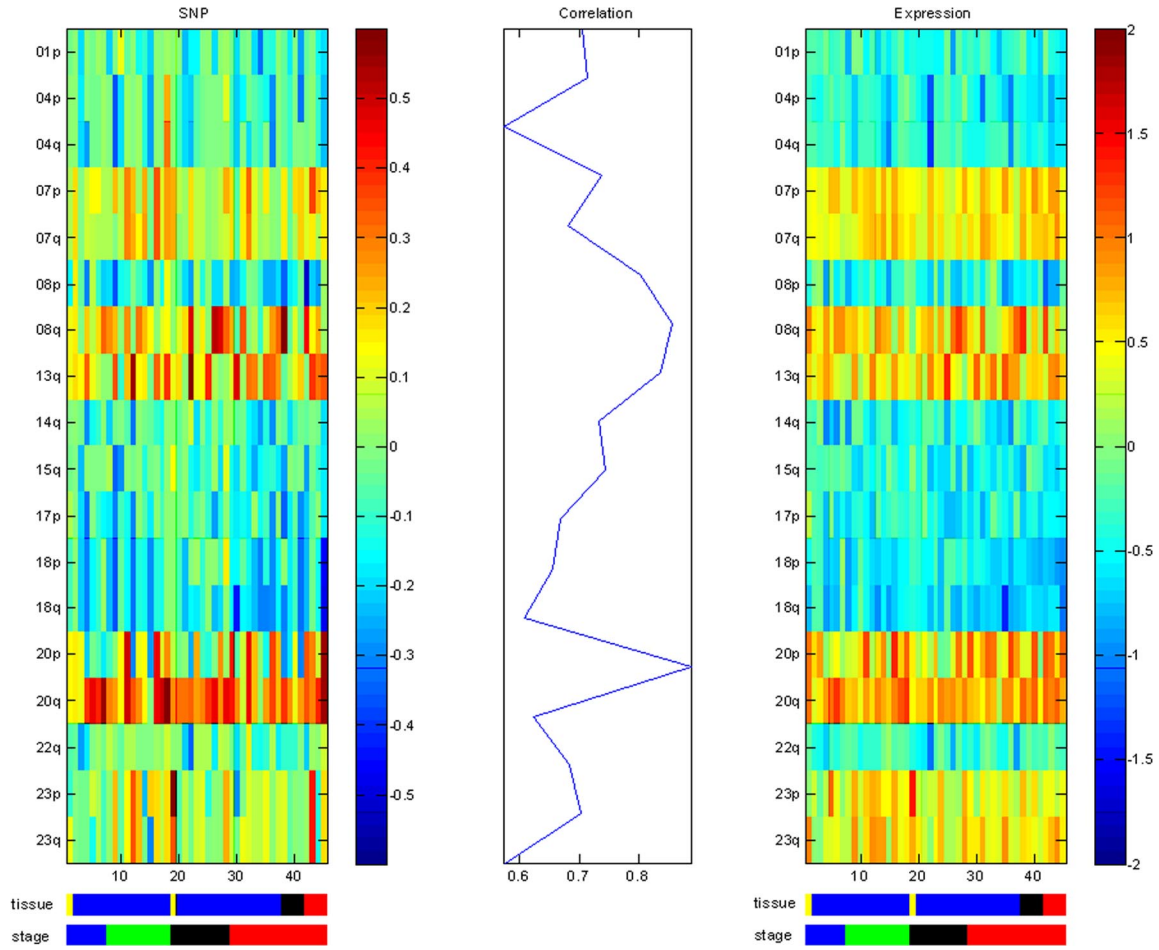

Fig. S1. Correlations between CINon copy number and expression tables. The *Left* (SNP) and *Right* (expression) matrices are deduced from the matrices shown in Fig. 2, showing only the 45 aneuploid samples for which we have both expression and copy number measures. The first color bar represents the tissue origin of the sample where primary tumor is blue, MIN tumor is yellow, lung metastasis is black, and liver metastasis is red. The second color bar stands for the stage of the tumor: blue, stage 1; green, stage 2; black, stage 3; red, stage 4 and metastasis. In the *Middle* is the plot of correlation between each pair of CINons. One can see that the correlations are high (above 0.6).

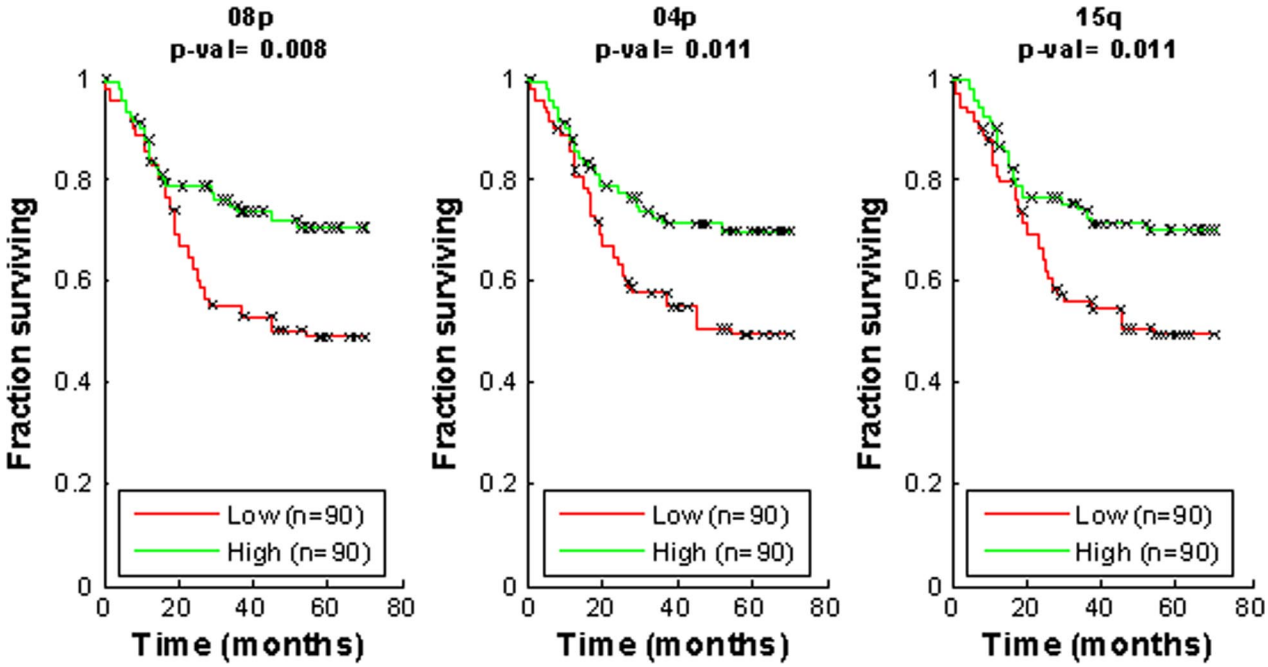

Fig. S2. Kaplan-Meier plots for separation on the basis of expression of 8p, 4p, and 15q, the chromosomal arms that passed 10% FDR. The *p*-values were calculated on the sorted CINon expression values (taken from the CINon expression table, see text) using 2 equal-sized high and low expression groups of the primary tumor samples. Follow-up intervals that were greater than 70 months were assigned 70.

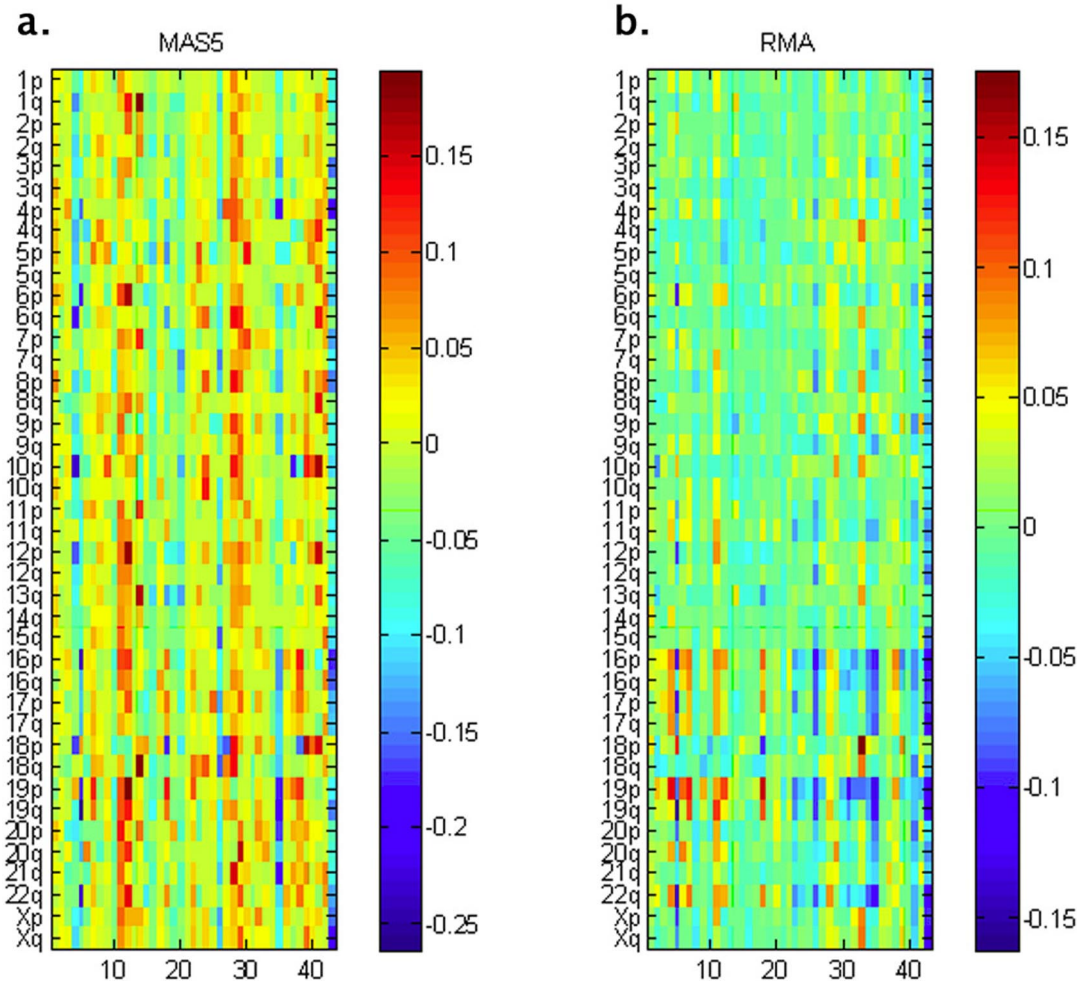

Fig. S3. MAS5 vs. RMA. The following matrices show the difference between MAS5 and RMA in the normal colon samples. The rows in each matrix represent chromosomal arms and the columns represent normal colon samples. Each entry ($M_{i,s}$) is calculated as the median of all probe sets $ER_{n,s}$ located on chromosomal arm *i* and sample *s*. The matrix in *a* represents the values after MAS5 normalization, as described in *SI Methods*. RMA values, as shown in *b* had the same preprocess stages, except they were not log2 transformed, and the fold-change values were calculated compared to all normal samples without taking into consideration the batch effect of the different protocols that were used (see *SI Methods*). For some reason RMA shows greater difference between chromosomes 1–15 and 16-X, that is not shown by the MAS5 algorithm.

Table S1. Table of correlated genes that reside on focal/peak CINons

Id	Symbol	CINon	Chr	Band	*Q*-value
222258_s_at	SH3BP4	1	2	q37.2	6.43*E*-15
202142_at	COPS8	1	2	q37.3	2.26*E*-05
221575_at	SCLY	1	2	q37.3	1.65*E*-07
221548_s_at	ILKAP	1	2	q37.3	1.37*E*-02
218301_at	RNPEPL1	1	2	q37.3	5.43*E*-01
210264_at	GPR35	1	2	q37.3	3.67*E*-06
218106_s_at	MRPS10	2	6	p21.1	4.74*E*-06
217931_at	CNPY3	2	6	p21.1	1.34*E*-10
218061_at	MEA1	2	6	p21.1	2.36*E*-10
214383_x_at	KLHDC3	2	6	p21.1	1.75*E*-01
36084_at	CUL7	2	6	p21.1	1.57*E*-10
207011_s_at	PTK7	2	6	p21.1	2.95*E*-12
202401_s_at	SRF	2	6	p21.1	7.28*E*-01
213204_at	PARC	2	6	p21.1	5.77*E*-01
39817_s_at	C6orf108	2	6	p21.1	1.25*E*-07
220554_at	SLC22A7	2	6	p21.1	7.47*E*-02
213485_s_at	ABCC10	2	6	p21.1	3.10*E*-12
47608_at	TJAP1	2	6	p21.1	2.05*E*-04
209317_at	POLR1C	2	6	p21.1	9.72*E*-23
219380_x_at	POLH	2	6	p21.1	9.89*E*-01
221050_s_at	GTPBP2	2	6	p21.1	1.17*E*-01
218385_at	MRPS18A	2	6	p21.1	3.60*E*-01
210512_s_at	VEGFA	2	6	p21.1	2.70*E*-24
201802_at	SLC29A1	2	6	p21.1	2.09*E*-16
200064_at	HSP90AB1	2	6	p21.1	7.79*E*-34
203927_at	NFKBIE	2	6	p21.1	1.78*E*-11
209056_s_at	CDC5L	2	6	p21.1	1.39*E*-08
203510_at	MET	5	7	q31.2	4.50*E*-30
207614_s_at	CUL1	6	7	q36.1	2.59*E*-07
203358_s_at	EZH2	6	7	q36.1	6.54*E*-10
202033_s_at	RB1CC1	9	8	q11.23	7.38*E*-03
221504_s_at	ATP6V1H	9	8	q11.23	8.33*E*-06
216241_s_at	TCEA1	9	8	q11.23	7.62*E*-08
212449_s_at	LYPLA1	9	8	q11.23	4.56*E*-05
218027_at	MRPL15	9	8	q11.23	1.52*E*-02
219231_at	TGS1	9	8	q12.1	1.85*E*-18
202625_at	LYN	9	8	q12.1	2.40*E*-08
214003_x_at	RPS20	9	8	q12.1	1.57*E*-15
205372_at	PLAG1	9	8	q12.1	5.43*E*-01
218642_s_at	CHCHD7	9	8	q12.1	3.99*E*-01
218516_s_at	IMPAD1	9	8	q12.1	2.43*E*-01
222087_at	PVT1	10	8	q24.21	2.35*E*-19
219299_at	TRMT12	10	8	q24.13	2.88*E*-02
209510_at	RNF139	10	8	q24.13	5.00*E*-01
209218_at	SQLE	10	8	q24.13	8.30*E*-17
201985_at	KIAA0196	10	8	q24.13	2.11*E*-02
202241_at	TRIB1	10	8	q24.13	7.91*E*-01
214532_x_at	POU5F1	10	8	q24.21	1.51*E*-06
202431_s_at	MYC	10	8	q24.21	1.21*E*-25
219099_at	C12orf5	11	12	p13.32	1.96*E*-06
204146_at	RAD51AP1	11	12	p13.32	6.32*E*-10
218258_at	POLR1D	13	13	q12.2	5.10*E*-19
204328_at	TMC6	16	17	q25.3	5.35*E*-06
201079_at	SYNGR2	16	17	q25.3	2.55*E*-08
202338_at	TK1	16	17	q25.3	2.45*E*-04
219394_at	PGS1	16	17	q25.3	5.00*E*-01
220370_s_at	USP36	16	17	q25.3	3.62*E*-03
206724_at	CBX4	16	17	q25.3	5.48*E*-26
222116_s_at	TBC1D16	16	17	q25.3	3.10*E*-17
217796_s_at	NPLOC4	16	17	q25.3	2.21*E*-02
210428_s_at	HGS	16	17	q25.3	2.13*E*-09
200654_at	P4HB	16	17	q25.3	6.59*E*-06
211716_x_at	ARHGDIA	16	17	q25.3	9.43*E*-01
204970_s_at	MAFG	16	17	q25.3	1.95*E*-03
202148_s_at	PYCR1	16	17	q25.3	3.53*E*-24

Id	Symbol	CINon	Chr	Band	*Q*-value
218908_at	ASPSCR1	16	17	q25.3	2.93*E*-12
212968_at	RFNG	16	17	q25.3	1.42*E*-03
217782_s_at	GPS1	16	17	q25.3	1.38*E*-04
64438_at	FLJ22222	16	17	q25.3	3.75*E*-01
218130_at	C17orf62	16	17	q25.3	8.34*E*-02
219862_s_at	NARF	16	17	q25.3	7.31*E*-11
203064_s_at	FOXK2	16	17	q25.3	8.81*E*-07
209076_s_at	WDR45L	16	17	q25.3	7.22*E*-12
208804_s_at	SFRS6	19	20	q13.11	3.95*E*-01
213837_at	L3MBTL	19	20	q13.11	5.88*E*-01
218709_s_at	IFT52	19	20	q13.12	1.84*E*-06
201710_at	MYBL2	19	20	q13.12	4.78*E*-19
209020_at	C20orf111	19	20	q13.12	4.26*E*-09
208429_x_at	HNF4A	19	20	q13.12	8.59*E*-01
219633_at	C20orf121	19	20	q13.12	1.77*E*-03
221471_at	SERINC3	19	20	q13.12	2.87*E*-05
217718_s_at	YWHAB	19	20	q13.12	6.00*E*-07
201870_at	TOMM34	19	20	q13.12	7.63*E*-26
205411_at	STK4	19	20	q13.12	2.38*E*-03
202071_at	SDC4	19	20	q13.12	1.30*E*-08
217770_at	PIGT	19	20	q13.12	1.22*E*-04
202954_at	UBE2C	19	20	q13.12	4.42*E*-30
205388_at	TNNC2	19	20	q13.12	3.52*E*-01
217592_at	ZSWIM1	19	20	q13.12	7.28*E*-01
202075_s_at	PLTP	19	20	q13.12	7.38*E*-03
89948_at	C20orf67	19	20	q13.12	4.16*E*-01
78330_at	ZNF335	19	20	q13.12	2.69*E*-02
203936_s_at	MMP9	19	20	q13.12	5.36*E*-14
219447_s_at	SLC35C2	19	20	q13.12	1.36*E*-11
55692_at	ELMO2	19	20	q13.12	2.98*E*-08
217875_s_at	PMEPA1	20	20	q13.31	9.22*E*-18
204092_s_at	AURKA	20	20	q13.2	9.17*E*-23
202190_at	CSTF1	20	20	q13.31	6.30*E*-09
217737_x_at	C20orf43	20	20	q13.31	4.62*E*-09
209590_at	BMP7	20	20	q13.31	2.31*E*-05
201558_at	RAE1	20	20	q13.31	3.07*E*-26
213405_at	RAB22A	20	20	q13.32	1.17*E*-04
202549_at	VAPB	20	20	q13.32	8.96*E*-03
221500_s_at	STX16	20	20	q13.32	2.49*E*-13
89476_r_at	NPEPL1	20	20	q13.32	3.95*E*-03
220607_x_at	TH1L	20	20	q13.32	2.12*E*-22
217801_at	ATP5E	20	20	q13.32	2.17*E*-07
217851_s_at	SLMO2	20	20	q13.32	2.27*E*-01
204554_at	PPP1R3D	20	20	q13.33	1.39*E*-04
211038_s_at	CROCCL1	24	1	p36.13	3.36*E*-03
201155_s_at	MFN2	24	1	p36.22	3.43*E*-16
212326_at	VPS13D	24	1	p36.22	7.79*E*-16
202481_at	DHRS3	24	1	p36.22	7.65*E*-03
217992_s_at	EFHD2	24	1	p36.21	1.62*E*-01
212146_at	PLEKHM2	24	1	p36.21	5.00*E*-01
218934_s_at	HSPB7	24	1	p36.13	6.53*E*-12
221813_at	FBXO42	24	1	p36.13	2.37*E*-01
202675_at	SDHB	24	1	p36.13	2.03*E*-10
209791_at	PADI2	24	1	p36.13	1.98*E*-31
221656_s_at	ARHGEF10L	24	1	p36.13	1.31*E*-01
212394_at	KIAA0090	24	1	p36.13	1.25*E*-01
216381_x_at	AKR7A3	24	1	p36.13	2.07*E*-11
202139_at	AKR7A2	24	1	p36.13	2.25*E*-13
37012_at	CAPZB	24	1	p36.13	7.53*E*-04
37005_at	NBL1	24	1	p36.13	6.80*E*-02
203649_s_at	PLA2G2A	24	1	p36.13	9.02*E*-06
218309_at	CAMK2N1	24	1	p36.12	3.30*E*-08
218246_at	C1orf166	24	1	p36.12	3.39*E*-10
209018_s_at	PINK1	24	1	p36.12	4.05*E*-11
201935_s_at	EIF4G3	24	1	p36.12	2.37*E*-10
201749_at	ECE1	24	1	p36.12	1.20*E*-05

Id	Symbol	CINon	Chr	Band	*Q*-value
203911_at	RAP1GAP	24	1	p36.12	1.30*E*-07
214230_at	CDC42	24	1	p36.12	1.15*E*-06
219103_at	DDEFL1	24	1	p36.12	6.68*E*-09
202292_x_at	LYPLA2	24	1	p36.11	5.06*E*-04
202528_at	GALE	24	1	p36.11	3.63*E*-01
202772_at	HMGCL	24	1	p36.11	6.10*E*-20
202838_at	FUCA1	24	1	p36.11	1.44*E*-37
217779_s_at	PNRC2	24	1	p36.11	1.44*E*-08
202553_s_at	SYF2	24	1	p36.11	3.35*E*-08
209007_s_at	C1orf63	24	1	p36.11	5.43*E*-01
217766_s_at	TMEM50A	24	1	p36.11	7.31*E*-11
57082_at	LDLRAP1	24	1	p36.11	3.42*E*-01
221269_s_at	SH3BGRL3	24	1	p36.11	5.96*E*-05
218547_at	DHDDS	24	1	p36.11	8.68*E*-30
208668_x_at	HMGN2	24	1	p36.11	8.38*E*-01
203379_at	RPS6KA1	24	1	p36.11	9.44*E*-15
212152_x_at	ARID1A	24	1	p36.11	9.03*E*-04
218799_at	ATPBD1B	24	1	p36.11	5.44*E*-03
209453_at	SLC9A1	24	1	p36.11	3.66*E*-18
219278_at	MAP3K6	24	1	p36.11	1.01*E*-01
212111_at	STX12	24	1	p35.3	1.34*E*-17
201756_at	RPA2	24	1	p35.3	9.92*E*-01
205309_at	SMPDL3B	24	1	p35.3	3.73*E*-09
218671_s_at	ATPIF1	24	1	p35.3	2.91*E*-06
219235_s_at	PHACTR4	24	1	p35.3	1.72*E*-12
218977_s_at	TRSPAP1	24	1	p35.3	1.21*E*-01
201696_at	SFRS4	24	1	p35.3	2.09*E*-10
202898_at	SDC3	24	1	p35.2	9.02*E*-03
204054_at	PTEN	29	10	q23.31	7.15*E*-07
211285_s_at	UBE3A	32	15	q11.2	6.24*E*-01
202604_x_at	ADAM10	33	15	q22.1	3.65*E*-01
217828_at	SLTM	33	15	q22.1	9.14*E*-10
218761_at	RNF111	33	15	q22.1	1.76*E*-01
209120_at	NR2F2	34	15	q26.2	3.39*E*-03
220180_at	CCDC68	39	18	q21.2	1.58*E*-30
218145_at	TRIB3	41	20	p13	2.83*E*-35
221827_at	RBCK1	41	20	p13	1.25*E*-17
212073_at	CSNK2A1	41	20	p13	1.85*E*-01
201052_s_at	PSMF1	41	20	p13	3.55*E*-06
219958_at	C20orf46	41	20	p13	8.49*E*-08
202897_at	SIRPA	41	20	p13	2.62*E*-02
208821_at	SNRPB	41	20	p13	4.15*E*-13
200875_s_at	NOL5A	41	20	p13	1.48*E*-24
203459_s_at	VPS16	41	20	p13	2.20*E*-03
213795_s_at	PTPRA	41	20	p13	1.73*E*-04
215544_s_at	UBOX5	41	20	p13	8.77*E*-08
204447_at	ProSAPiP1	41	20	p13	6.43*E*-15
218159_at	C20orf116	41	20	p13	3.44*E*-19
209171_at	ITPA	41	20	p13	1.25*E*-11
50314_i_at	C20orf27	41	20	p13	7.79*E*-16
212437_at	CENPB	41	20	p13	1.92*E*-01
201853_s_at	CDC25B	41	20	p13	1.36*E*-28
218809_at	PANK2	41	20	p13	2.14*E*-04
204668_at	RNF24	41	20	p13	2.97*E*-03
210357_s_at	SMOX	41	20	p13	4.23*E*-18
219570_at	C20orf23	43	20	p12.1	3.63*E*-01
217792_at	SNX5	44	20	p11.23	4.39*E*-01
37254_at	ZNF133	44	20	p11.23	1.64*E*-06
219951_s_at	C20orf12	44	20	p11.23	2.58*E*-02
205218_at	POLR3F	44	20	p11.23	2.35*E*-01
201582_at	SEC23B	44	20	p11.23	1.21*E*-01

The *q*-value for each probe set calculated from the *t*-test between primary tumors and normal samples.

Table S2. Expression levels of each putative TSG and oncogene separate the samples into 2 groups

	Pathway	Number of genes in the pathway
PMEPA1	Oxidative phosphorylation	26
	Valine leucine and isoleucine degradation	12
	Citrate cycle (TCA cycle)	8
	Glyoxylate and dicarboxylate metabolism	5
	Glutathione metabolism	8
	TACI and BCMA stimulation of B cell immune responses	5
	Electron Transport Reaction in Mitochondria	4
	Fatty acid metabolism	9
POLR1D	Leukocyte transendothelial migration	35
	B cell receptor signaling pathway	21
	Toll-like receptor signaling pathway	28
	Natural killer cell mediated cytotoxicity	32
	Ribosome	23
	T cell receptor signaling pathway	23
	Hematopoietic cell lineage	21
	Cytokine–cytokine receptor interaction	47
	Chondroitin sulfate biosynthesis	7
	Cell adhesion molecules (CAMs)	26
	Fc epsilon RI signaling pathway	18
CCDC68	Long-term depression	12
	Aminoacyl-tRNA biosynthesis	7
	p53 signaling pathway	10

The table details the pathway analysis for the genes that differentiate between the 2 groups of samples (see *SI Methods*). The separations into 2 groups was as follows: *PMEPA1*: 162 samples had overexpression of this gene and 48 samples had normal-like expression, 973 genes passed the *t*-test (5% FDR) between these 2 groups; *POLR1D*: 127 samples had overexpression and 83 samples were normal-like expression, 1,602 genes passed 5% FDR; *CCDC68*: 179 samples were underexpressed and 31 samples were normal-like expression, 802 passed 5% FDR.

Table S3. Summary of the samples used in the copy number and expression analysis

Tissue	Number of samples	Number of samples in expression	Number of samples in SNP			Number of samples in Both expression and SNP		
			All	Near-euploid	Aneuploid	All	Near-euploid	Aneuploid
Polyp	46	46	—	—	—	—	—	—
Normal colon	71	43	43	—	—	15	—	—
Normal liver	4	—	4	—	—	—	—	—
Normal lung	3	—	3	—	—	—	—	—
Primary tumor	187	180	62	22	40	55	18	37
Liver metastasis	24	21	8	1	7	5	1	4
Lung metastasis	15	9	10	2	8	4	—	4
Total	350	299	130	25	55	79	19	45
Primary tumor–stage I	28	28	11	4	7	11	4	7
Primary tumor–stage II	48	47	18	7	11	17	6	11
Primary tumor–stage III	50	49	14	4	10	13	3	10
Primary tumor–stage IV	61	56	19	7	12	14	5	9

Table S4. *T*-tests of chromosomal arms

Polyps vs. stage 1	Stage 1 vs. stage 2	Stage 2 vs. stage 3	Stage 3 vs. stage 2–4	Stage 1 vs. stages 2–4	Stage 1–2 vs. stages 3–4	Stage 1–3 vs. stage 4	Stage 4 vs. liver metastasis	Stage 4 vs. lung metastasis	WT p53 vs. mutant p53	WT kRAS vs. mutant kRAS	WT APC vs. mutant APC	Crohn vs. non-Crohn patients
14q	—	22q	—	15q	08p	04p	07q	—	20q *	20p	18p	13q
20q						08p	Xq		Xq			18p
20p							20q		13q *			
08q							07p		17p			
18q							Xp		07q			
15q							13q		Xp			
13q							18p		18q *			
04q									14q			
									20p *			
									07p			
									15q			
									04q			
									18p			
									04p			

For each separation of the samples into the 2 groups listed in the top row, we present those chromosomal arms whose broad CINons (including the focal CINon 1p) have copy numbers that differ significantly between the 2 groups listed. Association between copy number changes of the chromosomal arms marked by * were identified previously by Pincas et al. (unpublished data). These authors also found association of copy number changes in 20q, 13q, 7p, 17p, and 18q with *APC* mutation, which we have not observed, and no association of 20p with *kRAS* mutation. We also compared the mutant *p53* vs. wild—type *p53* in samples that showed expression of *p53* and received similar results (the first 12 of the 14 chromosomal arms passed FDR of 10%).

Table S5. Enriched pathways

	Ubiquitin mediated proteolysis	Epithelial cell signaling in Helicobacter pylori infection	Oxidative phosphorylation	Telomeres	Adherens junction	O-Glycan biosynthesis	Cell cycle	Ras Signaling Pathway	SNARE interactions in vesicular transport	Phospholipids as signalling intermediaries	Role of Erk5 in Neuronal Survival	Small cell lung cancer	Cyclin E Destruction Pathway	ADP-Ribosylation Factor	Cholera - Infection
Size	28	18	23	8	17	8	22	8	10	8	6	18	5	7	10
P-value	0.0002	0.0007	0.003	0.007	0.007	0.007	0.008	0.009	0.013	0.016	0.018	0.020	0.020	0.020	0.022
1	*UBE2D3*	*ATP6V1B2*	*ATP6V1B2*	*PPP2CB*	*IQGAP1*	*GALNT1*	*YWHAZ*	*RALBP1*	*VAMP3*	*GNB1*	*PLCG1*	*MYC*	*RB1*	*PACSIN2*	*ACTB*
2	*UBE2L3*	*ATP6V1F*	*ATP6V1F*	*MYC*	*ACTB*	*OGT*	*PCNA*	*ELK1*	*VAMP2*	*AKT1*	*RPS6KA1*	*RB1*	*E2F1*	*ARFGAP3*	*ARF5*
3	*UBA1*	*TJP1*	*COX6C*	*TNKS*	*TJP1*	*GALNT7*	*HDAC1*	*AKT1*	*BET1*	*RAC1*	*AKT1*	*BCL2*	*CUL1*	*ARFGEF1*	*ATP6V1F*
4	*CUL4A*	*ADAM10*	*NDUFA6*	*POLR2A*	*EP300*	*GALNT11*	*SMC1A*	*RAC1*	*STX8*	*PTK2*	*MAPK1*	*LAMA3*	*TFDP1*	*CENTD1*	*PLCG1*
5	*UBE3C*	*LYN*	*SDHB*	*RB1*	*SMAD4*	*C1GALT1C1*	*CDC25B*	*NFKB1*	*STX1A*	*MAPK1*	*MEF2A*	*E2F1*	*FBXW7*	*ARFGAP1*	*ATP6V1C1*
6	*CUL4B*	*PLCG1*	*ATP6V1C1*	*TERF1*	*LMO7*	*C1GALT1*	*EP300*	*BCL2L1*	*SNAP23*	*PLCB1*	*MAPK7*	*CCNE2*		*ARFGEF2*	*SEC61G*
7	*UBE4B*	*ATP6V1C1*	*NDUFV2*	*BCL2*	*PTPN1*	*GCNT3*	*SMAD4*	*CDC42*	*VTI1B*	*SRC*		*BIRC4*		*DDEF1*	*ATP6V1E1*
8	*UBE2K*	*MAP2K4*	*ATP5J2*	*AKT1*	*YES1*	*B4GALT5*	*SMAD2*	*RALA*	*STX12*	*ASAH1*		*AKT1*			*ATP6V1D*
9	*TCEB1*	*MET*	*NDUFC1*		*SMAD2*		*RB1*		*YKT6*			*PTK2*			*ARF6*
10	*UBE2C*	*ATP6AP1*	*COX5A*		*PTPRM*		*DBF4*		*STX16*			*NFKB1*			*ATP6V1H*
11	*BIRC4*	*RAC1*	*COX10*		*MET*		*E2F1*					*IKBKB*			
12	*CUL1*	*ATP6V1E1*	*NDUFB1*		*RAC1*		*ORC5L*					*LAMA5*			
13	*UBR5*	*ATP6V1D*	*ATP6AP1*		*CSNK2A1*		*CCNE2*					*COL4A2*			
14	*UBE2G1*	*NFKB1*	*ATP6V1E1*		*MAPK1*		*RBL1*					*COL4A1*			
15	*CDC16*	*IKBKB*	*ATP6V1D*		*SRC*		*CUL1*					*BCL2L1*			
16	*ITCH*	*SRC*	*ATP5I*		*CDC42*		*PRKDC*					*MAX*			
17	*UBE3A*	*CDC42*	*ATP5A1*		*SMAD3*		*MCM7*					*TRAF3*			
18	*NEDD4L*	*ATP6V1H*	*NDUFA4*				*CDC16*					*PIAS2*			
19	*WWP1*		*ATP5E*				*MCM4*								
20	*SMURF1*		*UCRC*				*TFDP1*								
21	*UBOX5*		*NDUFB11*				*YWHAB*								
22	*PPIL2*		*PPA2*				*SMAD3*								
23	*UBE2H*		*ATP6V1H*												
24	*HERC2*														
25	*HERC1*														
26	*UBE2W*														
27	*FBXW7*														
28	*PIAS2*														

List of pathways enriched, within the list of correlated genes that were found in all the broad CINons (including the focal CINon 1p), and the corresponding genes. In each pathway, the genes located on deletons are marked in blue and the genes located on amplicons in red. Genes that are marked in black are located on CINons that were significantly deleted and amplified. Enrichment was calculated using the DAVID database (12, 13), as detailed in *SI Methods*. The oxidative phosphorylation pathway was significant for the same analysis on the list of genes located on deleted CINons.