Supplemental files for 'A network-based, integrative approach to identify genes with aberrant co-methylation in colorectal cancer'

Supplemental Figures

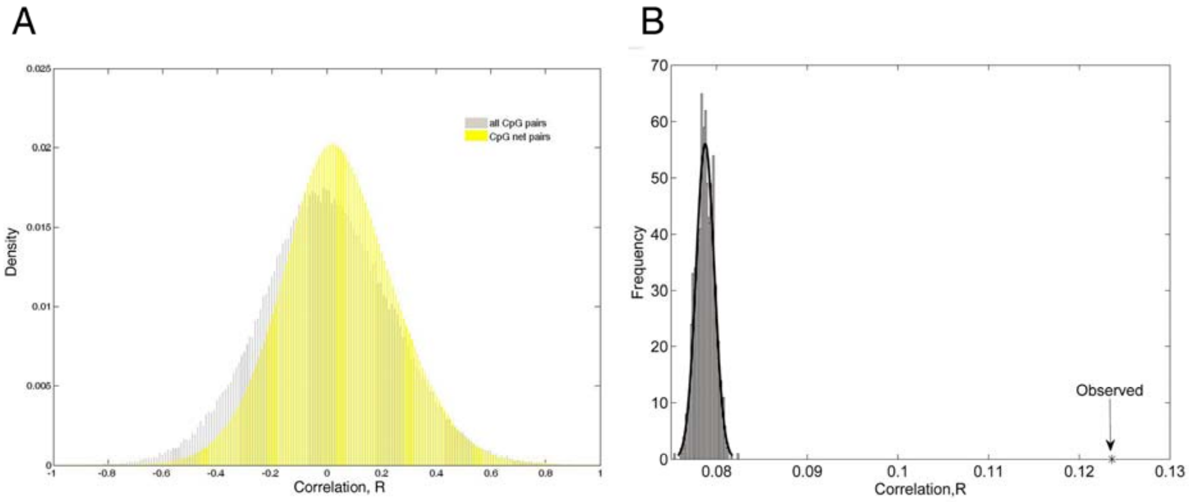

Figure S1. Connected CpG pairs in the functional network tend to be co-methylated.

(A) The distributions of Pearson correlation coefficient of methylation profiles for CpG connecting pairs and all CpG pairs in the colorectal cancer dataset. The gray bars were for all CpG pairs and yellow bars for CpG connecting pairs. **(B)** The mean Pearson correlation coefficient among connecting pairs is larger than random pairs. The average correlation of methylation is very low. The arrow represents the mean Pearson correlation coefficient of connecting pairs (average correlation is 0.124) in the colorectal cancer CpG connecting network, the line is fitted using random selected CpG pairs from the array (average correlation is 0.075).

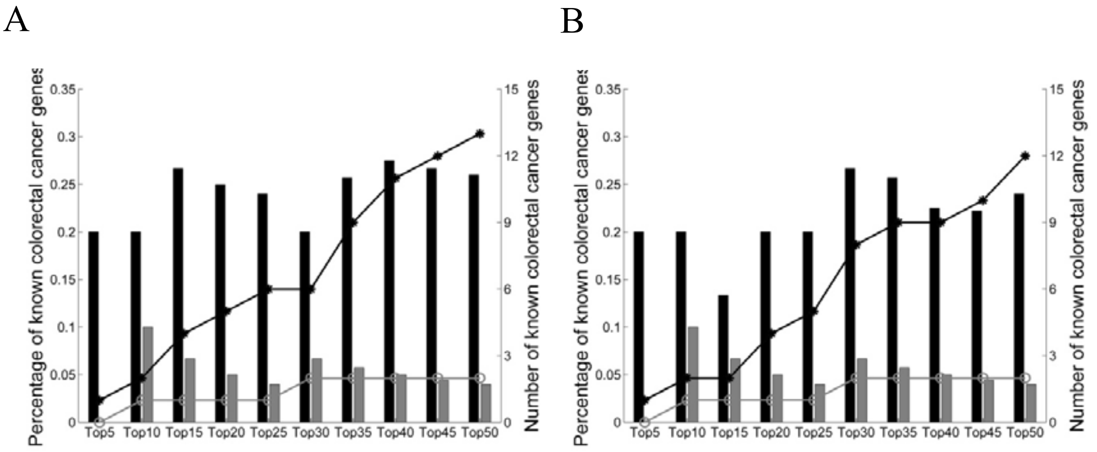

Figure S2. The robustness of the ICMAE for different *FDRs* with a grade of 5.

(A) The power of ICMAE to capture known colorectal cancer genes with *FDR*=0.05.
(B) The power of ICMAE to capture known colorectal cancer genes with FDR=0.15.

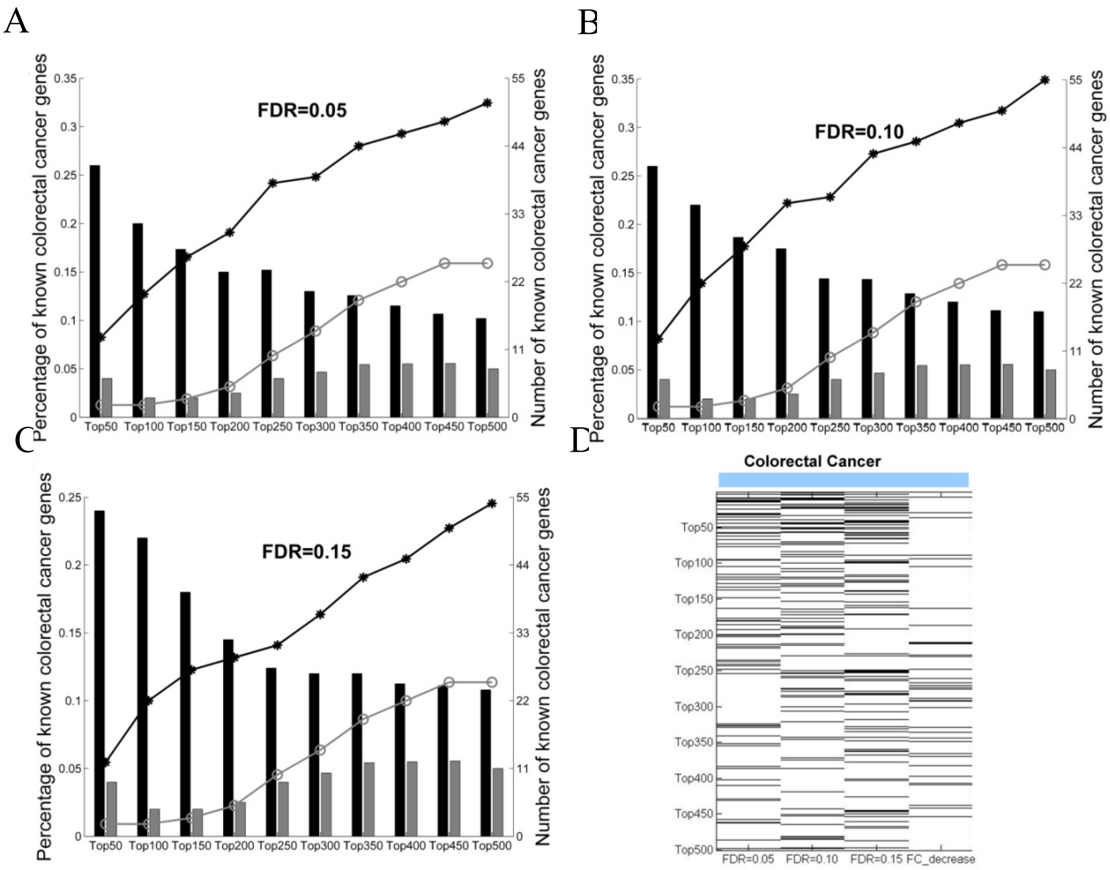

Figure S3. The robustness of the ICMAE for different *FDRs* with a grade of 50.

(A) The power of ICMAE and differential methylation with *FDR*=0.05. **(B)** The power of ICMAE and differential methylation with *FDR*=0.10. **(C)** The power of ICMAE and differential methylation with *FDR*=0.15. **(D)** The global view of the ability to capture known colorectal cancer genes.

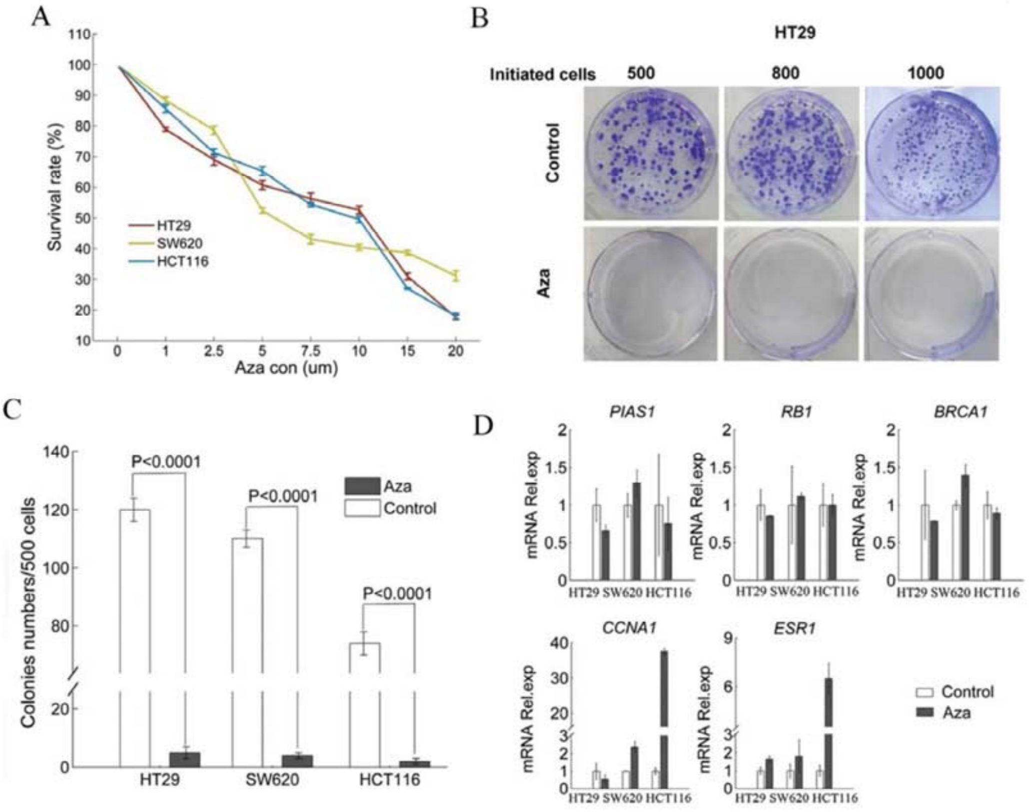

Figure S4. Function relevance of the potential colorectal cancer gene methylation.

(A) MTT analysis to determine the cell survival ability following 5-Aza treatment for three CRC cell lines. **(B)** and **(C)** Clonogenicity analysis of indicated CRC cell lines in the absence or presence of demethylation treatment with 5-Aza. **(D)** Relative expression (Rel.exp) of these five genes following demethylation treatment with 5-Aza in CRC cell lines. Confidence intervals represent standard deviations.

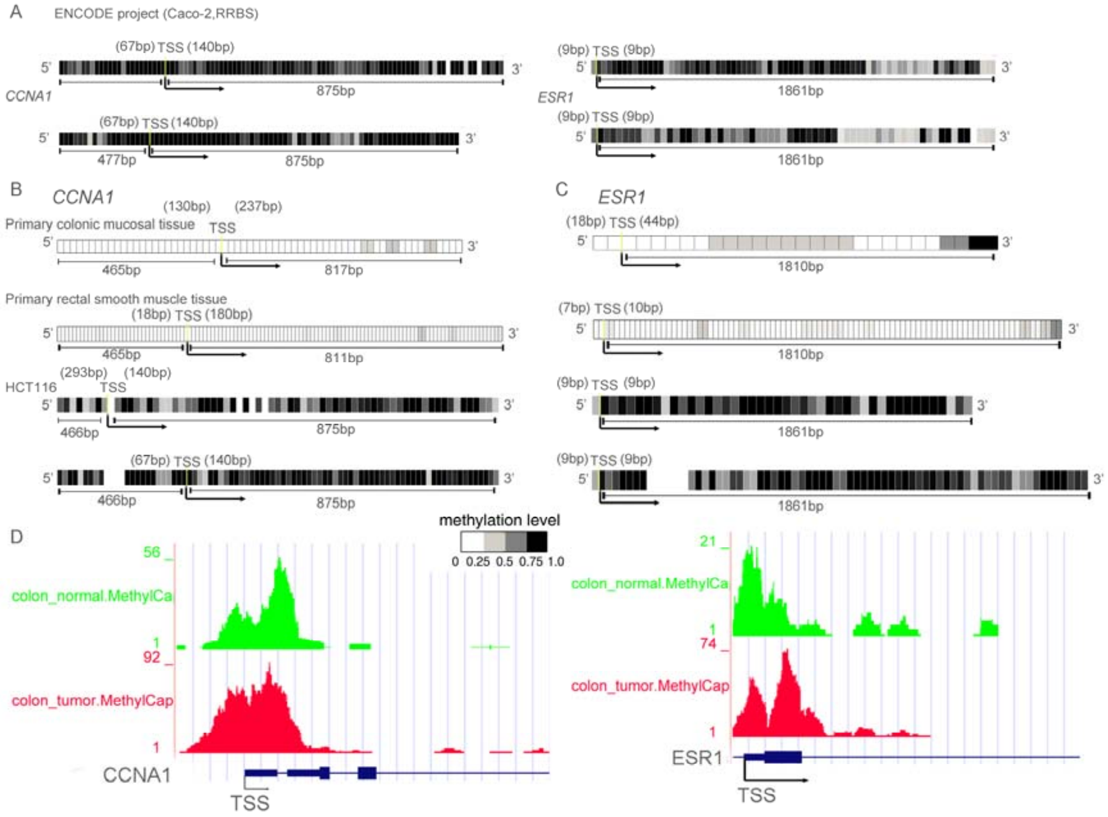

Figure S5. The methylation levels of CCNA1 and ESR1 in colorectal cancer.

(A) *CCNA1* and *ESR1* were highly methylated in Caco-2 cell line. Each block represents a CpG site located in the promoter and the methylation levels were indicated by the legend. The yellow lines mark the transcriptional start sites (TSS) and the numbers in bracket show the distances between TSS and adjacent CpG sites. We also marked the distance of the first and last CpGs to TSS. **(B)** and **(C)** *CCNA1* and *ESR1* were hypermethylated in HCT116 compared with primary colonic mucosal tissue or primary rectal smooth muscle tissue. **(D)** The genomic browser view shows that *CCNA1* and *ESR1* were hypermethylated in CRC. The height of the bar indicates the methylation levels measured by the number of methylated reads per 100bp bin.

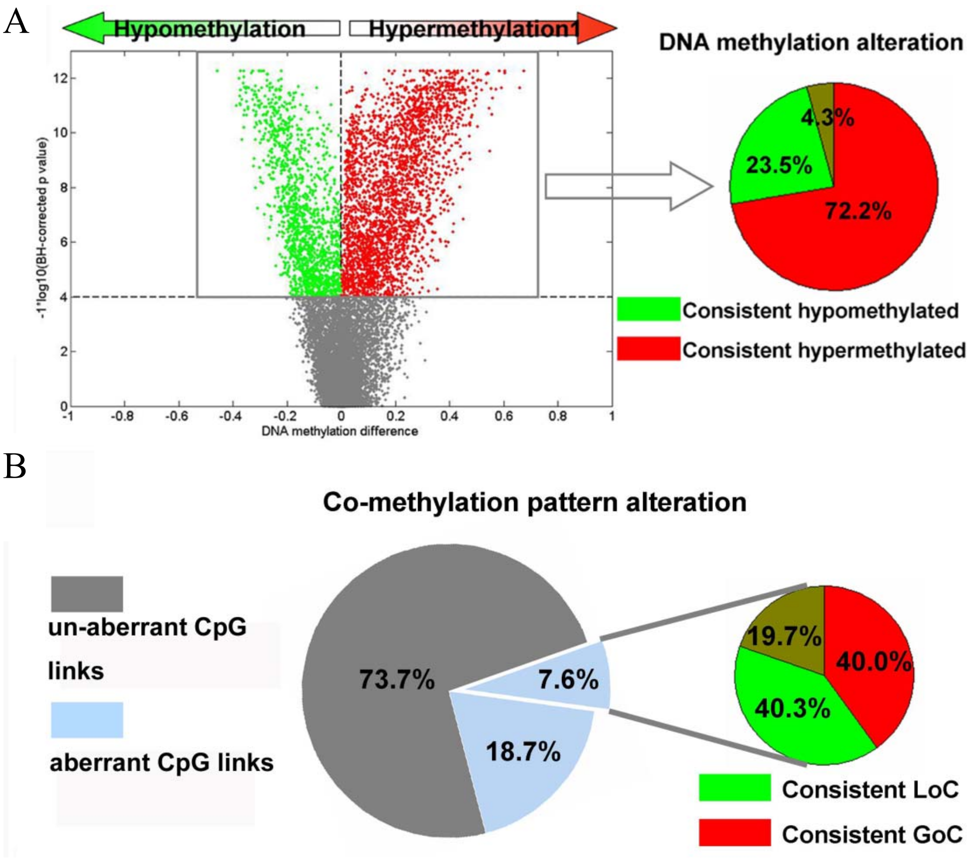

Figure S6. The consistency of aberration between interacting genes for colorectal cancer.

(A) The volcano plots of all CpG sites analyzed in the array. The beta value difference in the DNA methylation between tumor and normal samples is plotted on the *x* axis, and the *p* value for a *FDR*-corrected Wilcoxon rank test of differences (-1*$\log_{10}$scale) is plotted on the *y* axis. The pie plots the consistent DNA methylation alterations for genes with more than two CpG sites. **(B)** The consistent co-methylation alterations for aberrant interacting genes. Light blue pies represent the aberrant interactions and of which 7.6% gene pairs are with more than two CpG interaction pairs.

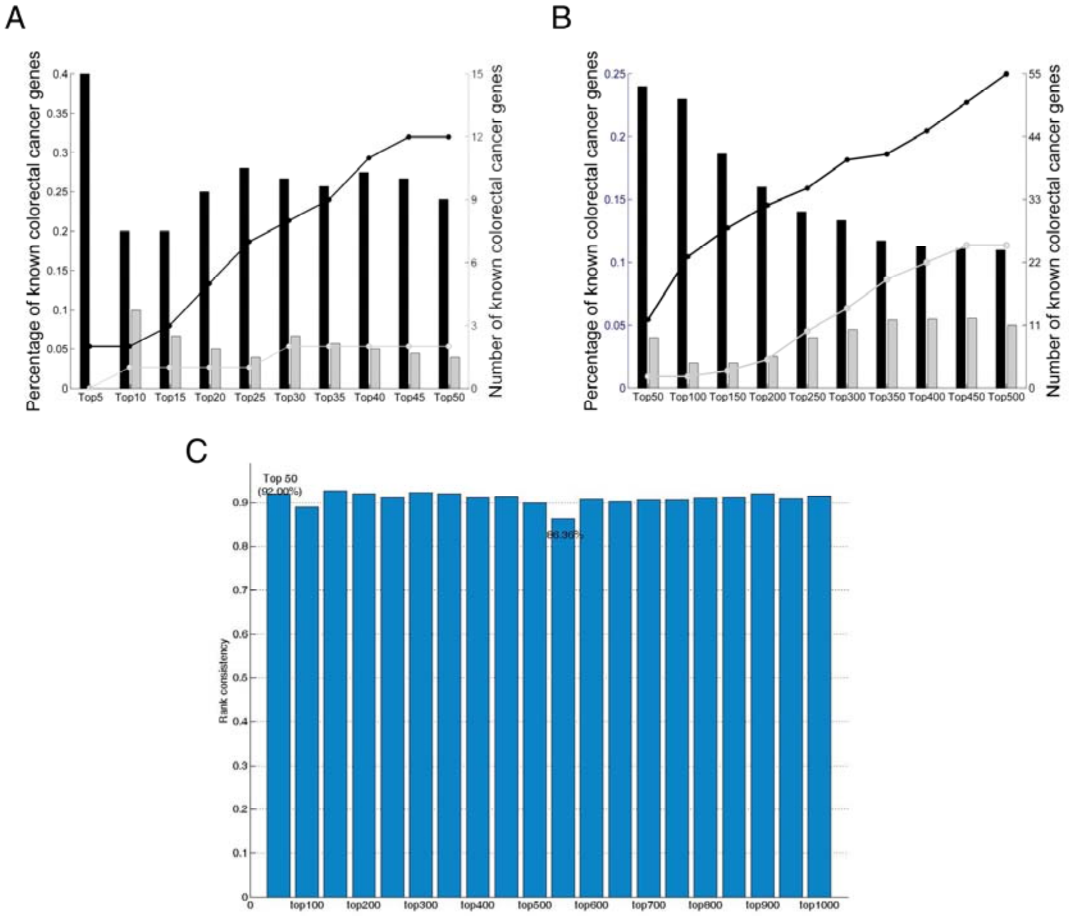

Figure S7. The ICMAE method is robust to outliers.

(A), The percentage of recalled known CRC genes in the top-ranked gene sets. The left y-axis represents the percentage of known CRC genes in the top-ranked gene set, corresponding to the bar figure. The right y-axis represents the number of recalled known CRC genes, corresponding to the line figure. Black is corresponding to genes ranked by their abnormal score computed by ICMAE strategy; grey is corresponding to differential methylated genes ranked by the degree of methylation changes. (B) The percentage of recalled known CRC genes in the top 500 gene sets. (C) The consistence of the ICMAE approach between processed outliers and non-processed outliers.

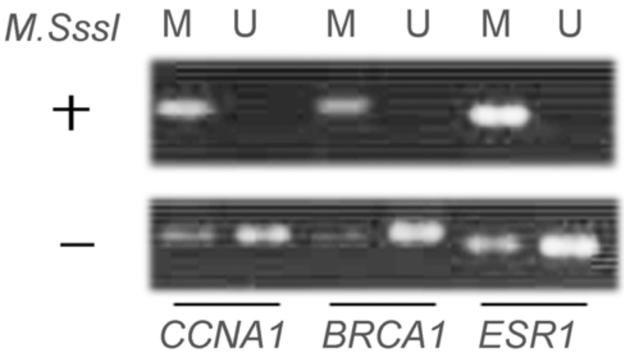

Figure S8. Validation of the specificity of the primers.

For each MSP reaction, we used placental DNA treated with M.SssI methytransferase or distilled water as positive and negative controls, respectively. The positive controls (+) were placental DNA treated with M.SssI methytransferase before bisulfite modification, and the negative control (-) reactions were the placental DNA without the treatment with M.SssI methytransferase before bisulfite modification.

Supplemental Text S1. Additional materials and methods.

Validation of the specificity of primers

DNA from the placental tissues was extracted by using the DNeasy Tissue Extraction Kit (Qiagen, Germany). For MSP reaction, we used placental DNA treated with M.SssI methyltransferase (New England Biolabs Inc, Beverly, Mass) or distilled water and then bisulfate modified as positive and negative controls. The Methylation specific PCRs were carried out with 30ng of bisulfate modified DNA in a total volume of 25μl, which contained 2.5μl 10×reaction buffer, 2.5mM MgCL2, 0.2mM dNTP, 10pmol forward and reverse primers, and one unit of HS Taq polymerse (Takara, Japan), in a S1000TM thermal cycler (Bio-Rad, USA). PCR reactions were denatured at 95℃ for 5 min, followed by 45 cycles of 30s at 95℃, 30s at the corresponding annealing temperature for each gene, 30s at 72℃, followed by a final extension step at 72℃ for 10 min. PCR products were visualized in a 2％ agarose gel stained with ethidium bromide at a final concentration of 0.1μg/ml. The results were shown in Fig. S8.

Drug treatment

Three cell lines (HCT116, SW620 and HT29), originally obtained from the American Type Culture Collection, were used for the study. These cell lines were all treated with 5-Aza-2′-deoxycytidine (5-Aza; Sigma, USA) for 96h at different concentrations indicated in the Figure S4 and then harvested on day 4. After overview the inhibiting effects of different concentrations and the results of previous studies[1, 2], we used the 5um to treat the cells and explored the expression of these genes. As the drug is very toxic, it might be expected that it would kill all cells, including non-transformed cells. David et al have demonstrated less toxicity to “normal” (non-CRC) cells[2], suggesting that the subsequent functional effects were due to de-methylation but not cell toxic.

Quantitative real-time RT-PCR

Total RNA was harvested with TRIzol Reagent (Invitrogen, Germany) following manufacturer’s instructions. cDNA was synthesized from 1μg of total RNA using

High Capacity cDNA reverse transcription Kit (Applied Biosystems, USA) and amplified by quantitative real-time RT-PCR with Fast Start Universal SYBR Green Master (Roche , Japan) using a step one Plus system (Applied Biosystems, USA). All primer sequences used for Real-time PCR are listed in Table S4.

MTT assays

Cell growth was measured by an MTT assay. Briefly, HCT116, HT29 and SW620 were seeded in 96-well plates and incubated at 37℃ for 12 h, and then treated with different concentrations of 5-Aza (sigma) or PBS. After 4 days, 100μg of tetrazolium-dye (3-(4, 5-dimethylthiazolyl-2)-2, 5-diphenyltetrazoliumbromide, MTT; Sigma) were added to each well and the cells were incubated for 4 h to reduce the dye. Next, the cells were treated with DMSO (Sigma) after which the absorbance of the converted dye in the living cells was measured using a microplate reader (BIO-RAD, Model 680, USA) at a wavelength of 490 nm. Six replicate wells were used for each analysis, and at least three independent experiments were conducted.

Colony formation assays

Three CRC cell lines were seeded at low density (500, 800 and 1000 cells/well) in a six-well plate and cultivated for 24h. Then the cells were treated with 5-Aza in concentration in medium containing 10% FCS for up to 15 days until countable clones had developed. Medium exchange was carried out regularly. Clones were stained with methylene blue (Merck, Darmstadt, Germany) for 10 minutes at room temperature. Then the plates were rinsed several times with distilled water and air-dried afterwards. The samples were tested in triplicates.

Supplemental Text S2. The *CCNA1* and *ESR1* were hypermethylated in CRC measured by BS-seq.

In order to measure the methylation of *CCNA1* and *ESR1* using a more quantitative method, we analyzed seven public BS-seq datasets. Two datasets of Caco-2 cell line and two of HCT-116 were obtained from Encyclopedia of DNA Elements (ENCODE) project[3]. The ENCODE consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. We directly downloaded the bed format from UCSC and obtained the methylation levels of CpGs located around these two genes. Due to difference in coverage, there are different numbers of CpGs in different datasets. Besides these colorectal cancer dataset, we also searched the Gene Expression Omnibus (GEO) and obtained two normal samples[4]. These datasets were processed using the software of Bismark to obtain the methylation levels of individual CpG site. Bismark is a tool to map bisulfite treated sequencing reads and performs methylation calling in a quick and easy-to-use fashion[5]. In addition, Brinkman et al. have provided the dataset of a paired colorectal tissue[6]. We downloaded the wiggle format and viewed the dataset in genomic browser view of UCSC. As a result, we found these two genes were highly methylated in CRCs (Figure S5).

Supplemental Text S3. The effect of outliers on the results of ICMAE.

In our study, we used Pearson correlation coefficient to measure the co-methylation levels of two CpG sites in the CpG connecting network. Pearson's correlation measures the strength of the association between two variables, which is the most commonly used measure of association in biomedical research[7, 8]. However, the technique is sensitive to outliers to some extent. In order to explore the effects of outliers to our results, we preprocessed the colorectal cancer dataset. Firstly, similar as the previous study[9], we identified the outliers using the 'boxplot' function. Secondly, we dealt with the outliers as missing values and imputed the missing values. And then we reanalyzed the dataset. We found the dataset we used in our study is high quality. About 98.82% of the CpG sites we analyzed were with a ratio of outliers less than 10%. Using the preprocessed dataset, we reconstructed the aberrant co-methylation network in CRC. As a result, we found that 93.93% of GoCs and 94.83% of LoCs were in line with our original analyses. Compared with differential methylation analysis, we found that the proposed ICMAE method also increase the power to identify known colorectal cancer genes (Figure S7A-B below). In addition, the rank consistencies were all over 86.36% for the top 1000 CpG sites (Figure S7C below). Specifically, of the top 50 CpG sites analyzed in our manuscript, we found that 92.00% were also ranked in top 50, the maximum rank of these CpG sites was 150th in our re-analyses, indicating the results of the proposed approach is robust. We reminded the users to pay attention to the outliers in using our proposed approach. We can preprocess the DNA methylation dataset and then used the proposed ICMAE approach to identify the aberrant methylated genes. We expected the power of ICMAE may be improved if we preprocess the dataset properly.

Supplemental Tables

Table S1. The known colorectal cancer genes collected from databases.

Table S2. List of oligonucleotides used for MSP sequencing of *ESR1*, *CCNA1* and *BRCA1* genes.

Genes	The sequences
ESR1	MF: 5´-AAGTGGGGTTGGAGATATTTAAC-3´ MR: 5´-ACCTTTAACAAAAACTTCACTCGAA-3´, UF: 5´-AGTGGGGTTGGAGATATATTTAATG-3´ UR: 5´-ACCTTTAACAAAAACTTCACTCAAA-3´,
CCNA1	MF: 5´- GTATAAAGGATTAGGTTTCGTGAGC-3´ MR: 5´- GCCACTATTAAATATCTTCCCGAA-3´, UF: 5´- ATAAAGGATTAGGTTTTGTGAGTGT-3´ UR: 5´- CACCACTATTAAATATCTTCCCAAA-3´,
BRCA1	MF: 5´- TATTTTTTCGTAAGTAGAGGGAGTC-3´ MR: 5´- ATTCAAAATACGAAATAACAACGTA-3´, UF: 5´- TTTTTTTGTAAGTAGAGGGAGTTGG-3´ UR: 5´- ATTCAAAATACAAAATAACAACATA-3´,

Annealing temperature for every set of primers:
MSP: *ESR1* UF and UR, MF and MR: 53.5℃； *CCNA1* UF and UR, 60.3℃; MF and MR: 55.9℃；
BRCA1 UF and UR, MF and MR: 52.7℃.

Table S3. Five-fold cross-validation results of the SVM classifier using the top ranked 50 CpG sites in colorectal cancer as a signature.

Dataset	Accuracy	Sensitivity	Specificity	AUC
Hinoue et al.	0.942±0.027	0.984±0.022	0.760±0.086	0.974
Kim et al.	0.858±0.195	1.000±0.000	0.860±0.129	0.967
TCGA	0.951±0.035	1.000±0.000	0.721±0.203	0.984
YH Kim et al.	0.898±0.056	0.941±0.047	0.856±0.081	0.822

Table S4. List of oligonucleotides used for real-time PCR of *ESR1, CCNA1, BRCA1, RB1 and PIAS1 genes*.

Genes	The sequences
ESR1-real time	Sense: 5´- ATCCACCTGATGGCCAAG -3´ Anti-sense: 5´- GCTCCATGCCTTTGTTACTCA-3´
RB1-real time	Sense: 5´- TCCTGAGGAGGACCCAGAG-3´ Anti-sense: 5´- AGGTTCTTCTGTTTCTTCAAACTCA-3´
BRCA1-real time	Sense: 5´- TTGTTGATGTGGAGGAGCAA-3´ Anti-sense: 5´- CAGATTCCAGGTAAGGGGTTC-3´
CCNA1-real time	Sense: 5´- TCAGTACCTTAGGGAAGCTGAAA-3´ Anti-sense: 5´- CCAGTCCACCAGAATCGTG-3´
PIAS1-real time	Sense: 5´- GGACCTGTCCTTCCCTATCTC-3´ Anti-sense: 5´- CTGGAGATGCTTGATGTGGA-3´

Supplemental references

1. M. Chen, J. Zhang, N. Li, Z. Qian, M. Zhu, Q. Li, J. Zheng, X. Wang and G. Shi, *PloS one*, 2011, **6**, e25564.
2. D. Mossman, K. T. Kim and R. J. Scott, *BMC cancer*, 2010, **10**, 366.
3. L. L. Elnitski, P. Shah, R. T. Moreland, L. Umayam, T. G. Wolfsberg and A. D. Baxevanis, *Genome Res*, 2007, **17**, 954-959.
4. B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen and J. A. Thomson, *Nat Biotechnol*, 2010, **28**, 1045-1048.
5. F. Krueger and S. R. Andrews, *Bioinformatics*, 2011, **27**, 1571-1572.
6. A. B. Brinkman, H. Gu, S. J. Bartels, Y. Zhang, F. Matarese, F. Simmer, H. Marks, C. Bock, A. Gnirke, A. Meissner and H. G. Stunnenberg, *Genome Res*, 2012, **22**, 1128-1138.
7. B. Zhang, C. Gaiteri, L. G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu and V. Emilsson, *Cell*, 2013, **153**, 707-720.
8. D. Kvitsiani, S. Ranade, B. Hangya, H. Taniguchi, J. Z. Huang and A. Kepecs, *Nature*, 2013, **498**, 363-366.
9. R. Akulenko and V. Helms, *Hum Mol Genet*, 2013, **22**, 3016-3022.