## Supplemental methods

### *BRAF* and *KRAS* mutation analysis

*BRAF* mutations at codon 600 in exon 15 and *KRAS* mutations at codons 12 and 13 in exon 2 were identified using the pyrosequencing assay on a Pyrosequencing 96HS (Biotage, Uppsala Sweden) per the manufacturer's protocol. A 224 bp fragment of the *BRAF* gene containing exon 15 was amplified from genomic DNA using the following primers: 5' TCA TAA TGC TTG CTC TGA TAG GA 3' and 5'Biotin-GGC CAA AAA TTT AAT CAG TGG A 3'and genotyped with the sequencing primer 5' CCA CTC CAT CGA GAT T 3'. Similarly, a 214 bp fragment of the *KRAS* gene containing exon 2 was amplified from each genomic DNA sample using the following primers: 5'Biotin-GTG TGA CAT GTT CTA ATA TAG TCA 3' and 5' GAA TGG TCC TGC ACC AGT AA 3' and genotyped with the sequencing primer 5' GCA CTC TTG CCT ACG 3'.

### *TP53* mutation analysis

*TP53* exons 4 through 8 were amplified by PCR using three exon-specific primer sets: Exon 4, 5'-GTT CTG GTA AGG ACA AGG GTT-3' (forward) and 5'-CCA GGC ATT GAA GTC TCA TG-3' (reverse) (Tm=49°C); Exons 5 and 6, 5'-GGT TGC AGG AGG TGC TTA C-3' (forward) and 5'-CCA CTG ACA ACC ACC CTT AAC-3' (reverse) (Tm=51°C); Exons 7 and 8, 5'-CCT GCT TGC CAC AGG TCT C-3' (forward) and 5'-TGA ATC TGA GGC ATA ACT GCA C-3' (reverse) (Tm=51°C). PCR amplification was performed using a touchdown protocol with an initial step of 95°C for 12 minutes, then 5 cycles of 95°C for 25 sec, Tm+15°C for 1 min and 72°C for 1 min, then 5 cycles of 95°C for 25 sec, Tm+10°C for 1 min and 72°C for 1 min, followed by 5 cycles of 95°C for 25 sec, Tm+5°C for 1 min and 72°C for 1 min, finishing with 35 cycles of 95°C for 25 sec, Tm°C for 1 min and 72°C for 1 min.

Sequencing of the purified PCR products was performed using an ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction Kit. Cycle sequencing reactions were performed in a thermal cycler for 25 cycles at 96°C for 10 sec, annealing at 50°C for 5 sec, and extension at 60°C for 4 min. Prior to capillary electrophoresis, unincorporated dye terminators were removed from the extension product using a DyeEx 96 Plate (Qiagen, Inc.) according to the manufacturer's instructions. The purified extension products were denatured at 90°C for 2 min and placed on ice for 1 min. Sequencing was performed on an ABI PRISM 3730xl DNA Analyzer (Applied Biosystems Inc.). The sequencing output files (.ab1) were processed using the Phred/Phrap software package developed at the University of Washington {Nickerson et al., 1997, Nucleic Acids Res, 25, 2745-51; Ewing et al., 1998, Genome research, 8, 175; Ewing and Green, 1998, Genome research, 8, 186; Gordon et al., 1998, Genome research, 8, 195}. Sequence Alignments for each exon read were viewed in the Consed Viewer Software and sequence variations were annotated and recorded. Samples containing missense mutations, nonsense mutations, splice-site mutations, frame-shift mutations, and in-frame deletions were considered positive for a mutation.

**MethyLight**

Genomic DNAs were treated with sodium bisulfite using the Zymo EZ DNA Methylation Kit (Zymo Research, Orange, CA) and subsequently analyzed by MethyLight as previously described {Campan et al., 2009, Methods Mol Biol, 507, 325-37}. The primer and probe sequences for the MethyLight reactions for the five gene CIMP marker panel and *MLH1* were reported previously {Weisenberger et al., 2006, Nat Genet, 38, 787-93}. The results of the MethyLight assays were scored as PMR (Percent of Methylated Reference) values as previously defined, with a PMR of ≥10 was used as a threshold for positive DNA methylation in each sample {Campan et al., 2009, Methods Mol Biol, 507, 325-37; Weisenberger et al., 2006, Nat Genet, 38, 787-93}. A sample was scored as CIMP-positive if ≥3 of the five CIMP-defining markers gave PMR values ≥10.

### Gene expression assay (Illumina HumanRef-8 v3.0 Expression BeadChip)

Total RNA from 25 pairs of colorectal tumor and non-tumor adjacent tissue samples was isolated using the TRIZOL® Reagent (Invitrogen, Burlington, ON) according to the manufacturer's protocol. The concentrations of RNA samples were measured using the NanoDrop 8000 (Thermo Fisher Scientific Inc., Waltham, MA). The quality of the RNA samples was assessed using the Experion RNA StdSens analysis kit (Bio-Rad, Hercules, CA). Expression analysis was performed using the Illumina Ref-8 whole-genome expression BeadChip (HumanRef-8 v3.0, 24,526 transcripts) (Illumina). Briefly, RNA samples were processed using the Illumina TotalPrep RNA Amplification Kit (Illumina). Total RNA (500ng) from each sample was subject to reverse transcription with an oligo(dT) primer bearing a T7 promoter. The cDNA then underwent second strand synthesis and purification. Biotinylated cRNA was then generated from the double-stranded cDNA template through *in vitro* transcription with T7 RNA polymerase. The biotinylated cRNA (750ng) from each patient was then hybridized to the BeadChips. The hybridized chips were stained and scanned using the Illumina HD BeadArray scanner (Illumina). Scanned image and bead-level data processing were performed using the BeadStudio 3.0.1 software (Illumina).

### Data masking for the Illumina Infinium DNA methylation assay

We masked data point as "NA" for probes that that contain single-nucleotide polymorphisms (SNPs) (dbSNP NCBI build 130/hg18) within the five base pairs from the interrogated CpG site or that overlap with a repetitive element that covers the targeted CpG dinucleotide. Furthermore, we replaced data points with "NA" for probes that are not uniquely aligned to the human genome (NCBI build 36/hg18) at 20 nucleotides at the 3' terminus of the probe sequence, and those that overlap with regions of insertions and deletions in the human genome. Together, we masked data points for 4,484 probes.

### Unsupervised consensus clustering

We applied the logit (logistic) transformation to DNA methylation β-values and median-centered each probe across the tumor samples. We then performed consensus clustering using the same 2,727 Infinium DNA methylation probes that were used for RPMM-based clustering. The optimal number of clusters was assessed based on 1,000 re-sampling iterations (seed value: 1022) of K-means clustering for K=2,3,4,5,6 with Pearson correlation as the distance metric as implemented in the R/Bioconductor *ConsensusClusterPlus* package.

### Validation of Infinium DNA methylation data by MethyLight assay

Genomic DNA from 25 pairs of colorectal tumor and their adjacent normal samples were treated with sodium bisulfite using the Zymo EZ96 DNA Methylation Kit (Zymo Research) and subsequently analyzed by MethyLight as previously described (Campan et al., 2009). Primers and probes used for validation are as follows and are listed as 5' to 3': *SFRP1,* forward primer: 5' GAA TTC GTT CGC GAG GGA 3', reverse primer: 5' AAA CGA ACC GCA CTC GTT

ACC 3’, probe: 6FAM-CCG TCA CCG ACG CGA AAA CCA AT-BHQ-1; *TMEFF2*, forward primer: 5’ GTT AAA TTC GCG TAT GAT TTC GAG A 3’, reverse primer: 5’ TTC CCG CGT CTC CGA C 3’, probe: 6FAM-AAC GAA CGA CCC TCT CGC TCC GAA-BHQ-1; *LMOD1*, forward primer: 5’ TTT TAA AGA TAA GGG GTT ACG TAA TGA G 3’, reverse primer: 5’ CCG AAC TAA CGA ATT CAC CGA C 3’, probe: 6FAM-TCG TCC CTA CTT ATC TAA CTC TCC GTA-MGBNFQ. The results of the MethyLight assays were scored as PMR (Percent of Methylated Reference) values as previously defined (Weisenberger et al., 2006; Campan et al., 2009).

**Validation of the Illumina gene expression array data by quantitative RT-PCR assay**

Total RNA sample from 25 pairs of colorectal tumor and non-tumor adjacent tissue samples were treated with DNase using DNA-*free*™ kit (Applied Biosystems, Foster City, CA, USA) to remove contaminating DNA. Reverse transcription reaction was performed using iScript Reverse Transcription Supermix for RT-PCR (Bio-Rad). Quantitative RT-PCR assays were performed with primers and probes obtained from Applied Biosystems (*SFRP1:* Hs00610060_m1_M; *TMEFF2*: Hs00249367_m1_M; *LMOD1:* Hs00201704_m1_M). The raw expression values were normalized to those of *HPRT1* (Hs99999909_m1_M).

## Supplemental Figure Legend

**Supplemental Figure 1.**

(*A*) Delta area plot showing the relative change in area under the consensus cumulative distribution function (CDF) curve (Monti et al., 2003). (*B*) Consensus matrix produced by *k-means* clustering ($K = 4$). (*C*) The heatmap representation of 125 colorectal tumor samples using the Infinium DNA methylation data as shown in Figure 1. Cluster membership of each sample derived from RPMM-based clustering and consensus clustering are indicated as vertical bars with distinct colors above the heatmap. (*D*) Contingency table comparing the cluster membership assignments between the two different clustering methods.

**Supplemental Figure 2.**

Histogram analysis of the number of methylated CIMP-defining MethyLight-based markers in colorectal cancer samples. (*A*) Histogram analysis of the number of CIMP loci methylated in all 125 colorectal tumor samples. (*B*) Histogram analysis of the number of CIMP-defining loci methylated in each RPMM-based tumor cluster membership identified in Figure 1.

**Supplemental Figure 3.**

Scatter plot analyses comparing DNA methylation profiles of colorectal tumor and adjacent-normal samples, stratified by their RPMM-based cluster membership.

**Supplemental Figure 4.**

Comparison of DNA methylation profiles between CIMP-H and CIMP-L tumors. (*A*) The volcano plot shows the $-1\times\log_{10}$ transformed FDR-adjusted *P*-value vs. the mean DNA methylation difference between CIMP-H and CIMP-L tumors. FDR-adjusted *P* = 0.001 and |Δβ| = 0.2 are used as a cutoff for differential methylation. Two CpG sites that are hypermethylated in CIMP-L tumors compared to CIMP-H tumors are indicated in green. (*B*) Heatmap representing Infinium DNA methylation β-values for the two CpG sites (labeled in green in panel A) that are significantly hypermethylated in CIMP-L compared with CIMP-H. The four DNA methylation-based subgroups are indicated above the heatmap. A color gradient from dark blue to yellow was used to represent the low and high DNA methylation β-values, respectively.

**Supplemental Figure 5.**

DNA structural and sequence characteristics associated with five different gene categories based on DNA methylation profiles in colorectal tumors. The five categories include: 1, CIMP-associated DNA methylation markers specific for the CIMP-H subgroup only; 2, CIMP-specific DNA methylation shared between both the CIMP-H and CIMP-L subgroups; 3, non-CIMP cancer-specific DNA methylation; 4, constitutively unmethylated across tumor and adjacent normal tissue samples; 5, constitutively methylated across tumor and adjacent normal tissue samples. Distribution of *(A)* observed CpG/expected CpG ratio and *(B)* GC content over 250 bp upstream and 250 bp downstream from the interrogated CpG dinucleotide on the Infinium DNA methylation BeadArray, (*C*) Distribution of the Takai and Jones-calculated CpG island length (Takai and Jones, 2002, Proc Natl Acad Sci U S A, 99, 3740-5), (*D, E*) Distribution of distances of Infinium DNA methylation probes to the nearest *(D) ALU* and *(E) LINE* repetitive element. In each box plot, the top and bottom edges are the 25th and 75th quartiles, respectively. The horizontal line within each box identifies the median. The whiskers above and below the box extend to at most 1.5 times the interquartile range (IQR).

**Supplemental Figure 6.**

Validation of the Infinium DNA methylation data and gene expression array data using MethyLight and quantitative RT-PCR (qRT-PCR), respectively. The validations were performed for three genes indicated above each scatter plot (*A*) Comparison of Infinium DNA methylation β-value (x-axis) and $\log_2$-transformed gene expression value from Illumina expression array (y-axis). (*B*) Validation of Infinium DNA methylation data by MethyLight technology. The x-axis represents Infinium DNA methylation β-value and the y-axis represents PMR value from MethyLight assay. Pearson correlation coefficients between the assays: 0.85 for *SFRP1*, 0.91 for *TMEFF2* and 0.96 for *LMOD1*. (*C*) Validation of Illumina expression array data by qRT-PCR assay. The x-axis represents $\log_2$-transformed array-based gene expression value and the y-axis represents $\log_2$-transformed relative copy number normalized to *HTPR1* using qRT-PCR assay. Pearson correlation coefficients between the gene expression platforms: 0.93 for *SFRP1*, 0.89 for *TMEFF2* and 0.91 for *LMOD1*. (*D*) Comparison of MethyLight PMR values (x-axis) and $\log_2$-transformed normalized relative copy number from qRT-PCR assay (y-axis). Black open circle: adjacent normal (n=25), red open circle: tumors in CIMP-L, Cluster3 and Cluster 4 (n=19), blue open circle: CIMP-H tumors (n=6).

**Supplemental Table 1.**

Genes that are constitutively methylated in normal samples, but show variable levels of DNA hypomethylation in tumors.

**Supplemental Table 2.**

List of probes that are significantly more methylated in both CIMP-H and CIMP-L tumors compared with Non-CIMP tumors.

**Supplemental Table 3.**

Twenty CpG sites associated with *KRAS* mutant tumors based on *P* value >0.05 and mean DNA methylation β-value difference >0.20

**Supplemental Table 4.**

Gene promoter classification among colorectal samples

**Supplemental Table 5.**

Genes that were hypermethylated with β-value difference >0.20 and showed more than 2-fold decrease in their gene expression levels in CIMP-H tumors