Genome-scale analysis of aberrant DNA methylation in colorectal cancer

Toshinori Hinoue,[1] Daniel J. Weisenberger,[1] Christopher P.E. Lange,[2,3] Hui Shen,[1] Hyang-Min Byun,[4] David Van Den Berg,[1] Simeen Malik,[1] Fei Pan,[1] Houtan Noushmehr,[1] Cornelis M. van Dijk,[5] Rob A.E.M. Tollenaar,[3] and Peter W. Laird[1,6]

[1]*Department of Surgery and Department of Biochemistry and Molecular Biology, University of Southern California, USC Epigenome Center, Los Angeles, California 90089-9601, USA;* [2]*Department of Surgery, Groene Hart Hospital, 2800 BB Gouda, The Netherlands;* [3]*Department of Surgery, Leiden University Medical Center, 2300 RC Leiden, The Netherlands;* [4]*Jane Anne Nohl Division of Hematology, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California 90033, USA;* [5]*Department of Pathology, Groene Hart Hospital, 2800 BB Gouda, The Netherlands*

Colorectal cancer (CRC) is a heterogeneous disease in which unique subtypes are characterized by distinct genetic and epigenetic alterations. Here we performed comprehensive genome-scale DNA methylation profiling of 125 colorectal tumors and 29 adjacent normal tissues. We identified four DNA methylation–based subgroups of CRC using model-based cluster analyses. Each subtype shows characteristic genetic and clinical features, indicating that they represent biologically distinct subgroups. A CIMP-high (CIMP-H) subgroup, which exhibits an exceptionally high frequency of cancer-specific DNA hypermethylation, is strongly associated with *MLH1* DNA hypermethylation and the *BRAF*^{V600E} mutation. A CIMP-low (CIMP-L) subgroup is enriched for *KRAS* mutations and characterized by DNA hypermethylation of a subset of CIMP-H-associated markers rather than a unique group of CpG islands. Non-CIMP tumors are separated into two distinct clusters. One non-CIMP subgroup is distinguished by a significantly higher frequency of *TP53* mutations and frequent occurrence in the distal colon, while the tumors that belong to the fourth group exhibit a low frequency of both cancer-specific DNA hypermethylation and gene mutations and are significantly enriched for rectal tumors. Furthermore, we identified 112 genes that were down-regulated more than twofold in CIMP-H tumors together with promoter DNA hypermethylation. These represent ~7% of genes that acquired promoter DNA methylation in CIMP-H tumors. Intriguingly, 48/112 genes were also transcriptionally down-regulated in non-CIMP subgroups, but this was not attributable to promoter DNA hypermethylation. Together, we identified four distinct DNA methylation subgroups of CRC and provided novel insight regarding the role of CIMP-specific DNA hypermethylation in gene silencing.

[Supplemental material is available for this article.]

Colorectal cancer (CRC) arises through the accumulation of multiple genetic and epigenetic changes. Somatic mutations in *APC*, *BRAF*, *KRAS*, *PIK3CA*, *TP53*, and other genes have been frequently observed in CRC and are considered to be drivers of colorectal tumorigenesis (Wood et al. 2007). In addition, the majority of sporadic CRCs (65%–70%) display chromosomal instability (CIN), characterized by aneuploidy, amplifications and deletions of subchromosomal genomic regions, and loss of heterozygosity (LOH) (Pino and Chung 2010).

Two major types of epigenetic modifications closely linked to CRC are DNA methylation and covalent histone modifications (Jones and Baylin 2007). Aberrant DNA methylation of CpG islands has been reported in the earliest detectable lesions in the colonic mucosa, aberrant crypt foci (ACF) (Chan et al. 2002). Promoter CpG island DNA hypermethylation is associated with transcriptional gene silencing and can cooperate with other genetic mechanisms to alter key signaling pathways critical to colorectal tumorigenesis (Baylin and Ohm 2006). A recent large-scale comparison between genes mutated and hypermethylated in CRC revealed significant overlap between these two alterations (Chan et al. 2008). Importantly, DNA hypermethylation appeared to be the preferred mechanism when a gene can be inactivated by either mutation or promoter DNA hypermethylation.

New insights into the mechanisms and the role of CpG island hypermethylation in cancer have emerged from recent studies using integrated analyses of the two types of epigenetic modifications. We and other groups have reported that genes that are targeted by Polycomb group (PcG) proteins in embryonic stem (ES) cells are susceptible to cancer-specific DNA hypermethylation (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007). PcG target genes are characterized by trimethylation of histone H3 lysine 27 (H3K27me3), are maintained in a low expression state, and are poised to be activated during development (Bernstein et al. 2007). More recently, it has been found that genes targeted by H3K27me3 in normal tissues acquire DNA methylation and lose the H3K27me3 mark in cancer (Gal-Yam et al. 2008; Rodriguez et al. 2008). Importantly, epigenetic switching of H3K27me3 and DNA methylation mainly occurs at genes that are not expressed in normal tissues. Furthermore, cancer-specific H3K27me3-mediated gene silencing has also been shown to inactivate tumor suppressor genes independent

[6]**Corresponding author.**
E-mail plaird@usc.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.117523.110.

of DNA hypermethylation in CRC (Jiang et al. 2008; Kondo et al. 2008).

Colorectal tumors with a CpG island methylator phenotype (CIMP) exhibit a high frequency of cancer-specific DNA hypermethylation at a subset of genomic loci and are highly enriched for an activating mutation of *BRAF* (*BRAF*^{V600E}) (Weisenberger et al. 2006). CRCs with CIN and CIMP have been shown to be inversely correlated (Goel et al. 2007; Cheng et al. 2008) and appear to develop in two separate pathways (Leggett and Whitehall 2010). DNA hypermethylation of some CIMP-associated gene promoters has been detected in early stages of colorectal tumorigenesis (Ibrahim et al. 2011). Furthermore, an extensive promoter DNA hypermethylation has been observed in the histologically normal colonic mucosa of patients predisposed to multiple serrated polyps, the proposed precursors of CIMP tumors (Young and Jass 2006). Notably, some of the distinct genetic and histopathological characteristics associated with CIMP tumors may be directly attributable to CIMP-mediated gene silencing. We reported that CIMP-associated DNA hypermethylation of *MLH1* is the dominant mechanism for the development of sporadic CRC with microsatellite instability (MSI) (Weisenberger et al. 2006). Furthermore, the CIMP-specific inactivation of *IGFBP7*-mediated senescence and apoptosis pathways may provide a permissive environment for the acquisition of *BRAF* mutations in CIMP-positive tumors (Hinoue et al. 2009; Suzuki et al. 2010).

Recent studies from several groups indicated that colorectal tumors with *KRAS* mutations may also be associated with a unique DNA methylation profile. CIMP-low (CIMP-L) tumors were originally shown to exhibit DNA hypermethylation of a reduced number of CIMP-defining loci (Ogino et al. 2006). CIMP-L was significantly associated with *KRAS* mutations, was observed more commonly in men than women, and appeared to be independent of MSI status. Shen et al. (2007) described the CIMP2 subgroup, which also showed DNA hypermethylation of CIMP-associated loci, but was highly correlated (92%) to *KRAS* mutations and not associated with MSI. A recent report from Yagi et al. (2010) reported the intermediate-methylation epigenotype (IME), which was also associated with *KRAS* mutations.

In light of these findings, there is confusion with regard to DNA methylation subtypes in CRC. It is not fully established whether CIMP-L, CIMP2, or IME represents a unique DNA methylation-based subgroup in CRC, as limited numbers of genomic regions were used to derive membership in these studies. Moreover, the types of genes targeted for DNA methylation in each subgroup and the effects of DNA hypermethylation on gene expression in each subtype have not yet been fully explored. To better characterize DNA methylation subgroups in CRC, we have performed comprehensive, genome-scale DNA methylation profiling of 125 primary colorectal tumors and 29 adjacent non-tumor colonic mucosa samples using the array-based Illumina Infinium HumanMethylation27 (HM27) Platform. We have also obtained gene expression data for the paired tumor and adjacent normal samples to assess the biological implications of DNA methylation-mediated gene silencing in CRC.

Results

DNA methylation–based colorectal cancer classification

We performed comprehensive genome-scale DNA methylation profiling of 125 colorectal tumor samples and 29 adjacent non-tumor colonic tissue samples using the Illumina Infinium HM27 DNA methylation assay, which assesses the DNA methylation status of 27,578 CpG sites located at the promoter regions of 14,495 protein-coding genes (Bibikova 2009). We also identified the mutation status of the *BRAF*, *KRAS*, and *TP53* genes in the tumor samples. We first determined CRC subtypes based on DNA methylation profiles in the collection of 125 tumor samples. We excluded probes that might be unreliable (see the Supplemental Methods) and probes that are designed for sequences on either the X or the Y chromosome. We chose the top 10% of probes with the highest DNA methylation variability based on standard deviation of the DNA methylation β-value across the entire colorectal tumor panel (2758 probes) and then performed unsupervised clustering using a recursively partitioned mixture model (RPMM). RPMM is a model-based unsupervised clustering method specifically developed for beta-distributed DNA methylation data such as obtained on the Infinium DNA methylation assay platform (Houseman et al. 2008). We identified four distinct tumor subgroups, indicated as clusters 1, 2, 3, and 4, by this approach (Fig. 1). The genetic and clinical features of each cluster are summarized in Table 1.

For comparison, we also performed resampling-based unsupervised consensus clustering (Monti et al. 2003) of the DNA methylation data set and also identified four DNA methylation-based clusters using this method. We compared the DNA methylation consensus cluster assignments for each sample to their RPMM-based cluster assignments and found substantial overlap with 80% (100/125) of the tumors showing agreement in cluster membership calls between these two different clustering methods (Supplemental Fig. 1). We based our subsequent analyses on cluster membership derived from the RPMM-based unsupervised clustering method. This method is well-suited for beta-distributed DNA measurements and has successfully identified DNA methylation profiles that are clinically relevant in normal and tumor samples from diverse tissues types (Christensen et al. 2009a,b, 2010, 2011; Marsit et al. 2009, 2011).

The cluster 1 subgroup is enriched for CIMP-positive colorectal tumors, as determined by the CIMP-specific MethyLight five-marker panel developed previously in our laboratory (*CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3*, *SOCS1*), as well as *MLH1* DNA hypermethylation using MethyLight technology (Fig. 1; Weisenberger et al. 2006). All of the tumors with a *BRAF* mutation belong to this subgroup, and nearly half of the tumors in this subgroup that do not harbor *BRAF* mutations carry mutant *KRAS* (Fig. 1). This subgroup is also characterized by a low frequency of *TP53* mutations (11%). Clinically, the majority of these tumors were found in female patients (71%) and a proximal location in the colon (86%), both of which have been previously found to be associated with CIMP-positive CRC defined by the MethyLight five-marker panel (Weisenberger et al. 2006).

Previous studies with a limited number of DNA methylation markers from several groups indicated the existence of additional DNA methylation-based subtypes in CRC that are associated with *KRAS* mutations. These subgroups have been variously described as CIMP-low (Ogino et al. 2006), CIMP2 (Shen et al. 2007), and Intermediate-methylation epigenotype (IME) (Yagi et al. 2010). It is not clear whether these classifications represent the same tumor subgroup or different subgroups within CRC. We found that although *KRAS* mutant tumors are represented across the four classes, they are more common in the cluster 2 subgroup compared with the other clusters (Fig. 1; Table 1). Interestingly, the proportion of the tumors that show DNA methylation at one or two loci of the MethyLight-based five-marker panel is substantially higher in the cluster 2 subgroup (62%) than in the cluster 3 (11%) or cluster 4 tumors (13%) (Supplemental Fig. 2). These genetic and

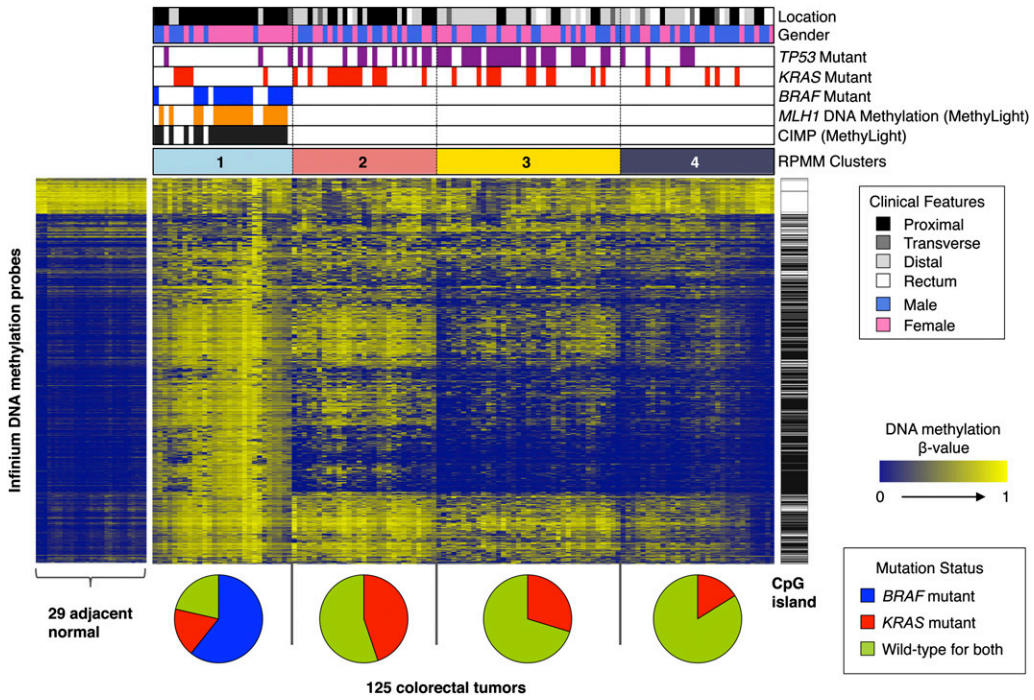

Figure 1. RPMM-based classification of 125 colorectal tumor samples and heatmap representation of Infinium DNA methylation data. DNA methylation profiles of 1401 probes with most variable DNA methylation values (standard deviation >0.20) in the 125 colorectal tumor sample set. The DNA methylation β-values are represented by using a color scale from dark blue (low DNA methylation) to yellow (high DNA methylation). Four subgroups were derived by RPMM-based clustering and are indicated *above* the heatmap: (light sky blue) cluster 1 ($n = 28$); (light coral) cluster 2 ($n = 29$); (yellow) cluster 3 ($n = 37$); (dark gray) cluster 4 ($n = 31$). (Black bars) CIMP-positive tumors as classified by the MethyLight five-marker panel (Weisenberger et al. 2006). Presence of *MLH1* DNA methylation (orange bars), *BRAF* mutation (blue bars), *KRAS* mutation (red bars), and *TP53* mutations (purple bars). Probes that are located within CpG islands (Takai-Jones) (Takai and Jones 2002) are indicated by the horizontal black bars to the *right* of the heatmap. The probes are arranged based on the order of unsupervised hierarchal cluster analysis using a correlation distance metric and average linkage method. Pie charts *below* the heatmap show the proportion of tumor samples harboring *BRAF* mutations (blue), *KRAS* mutations (red), and those that are wild-type for both *BRAF* and *KRAS* (yellow-green) within each subgroup.

epigenetic characteristics observed in the cluster 2 subgroup are consistent with the CIMP-low subtype described previously (Ogino et al. 2006). Therefore, in this study, we refer to the tumors that belong to the cluster 1 subgroup as CIMP-high (CIMP-H) and the cluster 2 subgroup tumors as CIMP-low (CIMP-L).

Our RPMM-based clustering analysis identified two other CRC subtypes, designated as clusters 3 and 4, in addition to the CIMP-H and CIMP-L subgroups (Fig. 1; Table 1). The difference between these two subgroups is not apparent based on DNA hypermethylation at CIMP-defining five-gene loci (Supplemental Fig. 2), indicating that DNA methylation signatures unrelated to CIMP might discriminate between these two CRC subsets. The frequency and level of cancer-specific DNA hypermethylation in the tumors in the cluster 4 subgroup appear to be the lowest among the four subclasses (Supplemental Fig. 3). Importantly, the tumors included in cluster 3 are distinguished by a significantly higher frequency of *TP53* mutations (65%; $P = 6.5 \times 10^{-5}$ [vs. cluster 4], Fisher's exact test) and their location in the distal colon (65%; $P = 0.028$ [vs. cluster 4], Fisher's exact test). In contrast, the tumors that belong to cluster 4 exhibit a lower frequency of both *KRAS* (16%) and *TP53* (16%) mutations, and their occurrence shows significant enrichment in the rectum compared with all the other groups ($P = 2.1 \times 10^{-3}$, Fisher's exact test). Cluster 4 tumors also show borderline statistical significance to be more commonly found in males compared to the cluster 3 tumors ($P = 0.056$, Fisher's exact test), providing additional lines of evidence that cluster 3 and cluster 4 tumors are distinct.

We also identified a panel of 119 gene promoters that are constitutively methylated in normal samples but show variable levels of DNA methylation in tumors (Fig. 1; for the list of genes, see Supplemental Table 1). It has long been established that the human genome is comprised primarily of sequences of low CpG density that are usually highly methylated in normal somatic tissues and that undergo loss of DNA methylation in cancer (Feinberg and Vogelstein 1983; Gama-Sosa et al. 1983; Miranda and Jones 2007). We found that, indeed, the majority of these probes are targeted to low-CpG density regions. The variable loss of DNA methylation among our tumor clusters is consistent with earlier reports that the degree of global DNA hypomethylation can vary considerably among colorectal tumors (Estecio et al. 2007). We performed a gene set enrichment analysis (GSEA) on these 119 genes using the Database for Annotation, Visualization and Integrated Discovery tool (DAVID). We found enrichment of genes involved in secretion (3.1-fold enrichment, $P = 1.9 \times 10^{-6}$), signaling (2.2-fold enrichment, $P = 6.8 \times 10^{-6}$), signal peptide (2.2-fold enrichment, $P = 2.5 \times 10^{-5}$), disulfide bond (2.3-fold enrichment, $P = 1.8 \times 10^{-5}$), and extracellular regions (2.3-fold enrichment, $P = 6.8 \times 10^{-4}$).

Characterization of the CIMP-H and CIMP-L subgroups

We next sought to investigate DNA methylation markers associated with the CIMP-H and CIMP-L subgroups. To accomplish this,

Table 1. Genetic and clinical features found in each of the four DNA methylation-based subtypes

		Overall		Cluster 1 (CIMP-H)		Cluster 2 (CIMP-L)		Cluster 3		Cluster 4	
Variable		*n*	%	*N*	%	*n*	%	*n*	%	*n*	%
Total		125	100	28	22	29	23	37	30	31	25
Gender	Female	65	52	20	71	12	41	22	59	11	35
	Male	60	48	8	29	17	59	15	41	20	65
Subsite	Proximal	54	43	24	86	15	52	7	19	8	26
	Transverse	7	6	1	4	1	3	2	5	3	10
	Distal	49	39	3	11	11	38	24	65	11	36
	Rectum	15	12	0	0	2	7	4	11	9	29
Stage	1 or 2	50	50	9	41	16	66	12	41	13	52
	3 or 4	50	50	13	59	8	34	17	59	12	48
	No info	25									
BRAF mutation	Mutant	17	14	17	61	0	0	0	0	0	0
	Wild-type	108	86	11	39	29	100	37	100	31	100
KRAS mutation	Mutant	34	27	5	18	13	45	11	30	5	16
	Wild-type	91	73	23	82	16	55	26	70	26	84
TP53 mutation	Mutant	43	34	3	11	11	38	24	65	5	16
	Wild-type	82	66	25	89	18	62	13	35	26	84
Age	Median	68		71		70		65		69	
	Range	33–90		51–90		33–87		44–88		34–87	
	No info	25									

we compared the DNA methylation β-values for each probe between CIMP-H and non-CIMP tumors (clusters 3 and 4 combined) as well as the β-values between CIMP-L and non-CIMP tumors using the Wilcoxon rank-sum test. We identified 1618 CpG sites that showed significant DNA hypermethylation in CIMP-H versus non-CIMP tumors (FDR-adjusted $P < 0.0001$) (Fig. 2A). In contrast, we found 435 CpG sites that are significantly hypermethylated in CIMP-L tumors compared with non-CIMP tumors (FDR-adjusted $P < 0.0001$) (Fig. 2A).

We observed substantial overlap between the CIMP-H- and CIMP-L-associated markers, as these appear to exhibit a higher frequency of promoter DNA hypermethylation in both tumor subgroups compared with non-CIMP tumors (Fig. 2A). Interestingly, we found that nearly 20% of CIMP-H-associated CpG sites (318 CpGs) are also methylated in CIMP-L tumors (FDR-adjusted $P < 0.0001$ vs. non-CIMP) (see list of genes in Supplemental Table 2).

To determine whether there are DNA methylation markers specifically associated with the CIMP-L subgroup, we examined 22 CpG sites that showed significant DNA hypermethylation in CIMP-L tumors but not in CIMP-H tumors, as compared to non-CIMP tumors (FDR-adjusted $P < 0.001$ [CIMP-L vs. non-CIMP] and $P > 0.05$ [CIMP-H vs. non-CIMP]) (Fig. 2A). Although these markers exhibited statistically significant DNA methylation differences, they did not show strong CIMP-L specificity when visualized and compared with individual tumor samples using a heatmap (Fig. 2B). We also directly compared the DNA methylation levels of each CpG locus between CIMP-H tumors and CIMP-L tumors (Supplemental Fig. 4A). We identified two CpG loci in the promoter regions of *SRRM2* and *NTF3* that are significantly hypermethylated in CIMP-L tumors compared with CIMP-H tumors ($P < 0.001$ and mean β-value difference > 0.2). Interestingly however, these two gene loci exhibit CIMP-H-specific DNA hypomethylation, as these are methylated in adjacent non-tumor colonic tissues, as well as in tumors that belong to the cluster 3 and cluster 4 subgroups (Supplemental Fig. 4B).

Specifically, we also did not find a significant increase in *MGMT* DNA hypermethylation in CIMP-L tumors compared with non-CIMP tumors ($P > 0.05$), as reported previously (Ogino et al. 2007). Clinically, Ogino et al. (2006) observed a significant association between CIMP-L and male sex. We also found that CIMP-L tumors are slightly more common in men (59%) than in women (41%), although the association did not achieve statistical significance ($P > 0.05$, Fisher's exact test).

Analysis of DNA methylation associated with *KRAS* mutant tumors

Significant enrichment of *KRAS* mutations in CIMP-L may suggest that *KRAS* mutations either induce DNA hypermethylation of a group of CpG loci or they might synergize with a specific DNA methylation profile associated with CIMP-L tumors. Interestingly, Shen et al. (2007) proposed a CIMP2 subtype of CRC, found to be tightly linked with *KRAS* mutations (92% of cases), using a limited number of DNA methylation markers.

We investigated whether *KRAS* mutations themselves are associated with DNA hypermethylation of specific sets of genes in CRC. We stratified tumors into three groups by their *BRAF* and *KRAS* mutation status: (1) *BRAF* mutant ($n = 17$); (2) *KRAS* mutant ($n = 34$); and (3) wild-type for both *BRAF* and *KRAS* ($n = 74$); and then compared DNA methylation profiles between each group. We identified a large number of CpG sites (715; FDR-adjusted $P < 0.0001$) that are significantly hypermethylated in tumors with *BRAF* mutation, all of which belong to the CIMP-H subgroup, as compared with tumors that are wild-type for *BRAF* and *KRAS* (Fig. 2C). In contrast, only one CpG locus located in the promoter of *JPH3* showed DNA hypermethylation in the *KRAS* mutant tumors compared to the *BRAF*/*KRAS* wild-type tumors at the 0.01 significance level (Fig. 2C). Using a less stringent significance threshold (FDR-adjusted $P < 0.05$), we identified 157 CpGs that showed more frequent DNA methylation in *KRAS* mutant tumors (Fig. 2C). However, we found that the mean β-value differences for the majority of these probes between tumors with a *KRAS* mutation and those that are wild-type for *BRAF*/*KRAS* are small (0.08 ± 0.09, mean $|\Delta\beta| \pm$ SD). Among the 157 probes, we further examined the 22 CpG sites that showed substantial mean β-value difference ($(|\Delta\beta| > 0.20)$ between *KRAS* mutant tumors and *BRAF*/*KRAS* wild-type tumors. Importantly, we found that all of these CpG sites exhibit CIMP-L-specific DNA hypermethylation with much higher significance levels (Wilcoxon rank-sum test between CIMP-L and non-CIMP tumors) (Supplemental Table 3). These observations indicate that the significant association between DNA methylation at these loci and the *KRAS* mutation is mainly due to CIMP-L-based DNA hypermethylation.

To further examine the DNA methylation profiles in *KRAS* mutant tumors and *BRAF*/*KRAS* wild-type tumors, we subdivided CIMP-L and non-CIMP tumors by their *KRAS* mutation status and compared the mean DNA methylation β-values among these groups. We observed that mean DNA methylation β-values for *KRAS* mutant tumors and those *BRAF/KRAS* wild-type tumors are well-correlated within both the CIMP-L and non-CIMP subgroups (Fig. 3A,B). Moreover, the CIMP-L subgroup exhibits higher mean

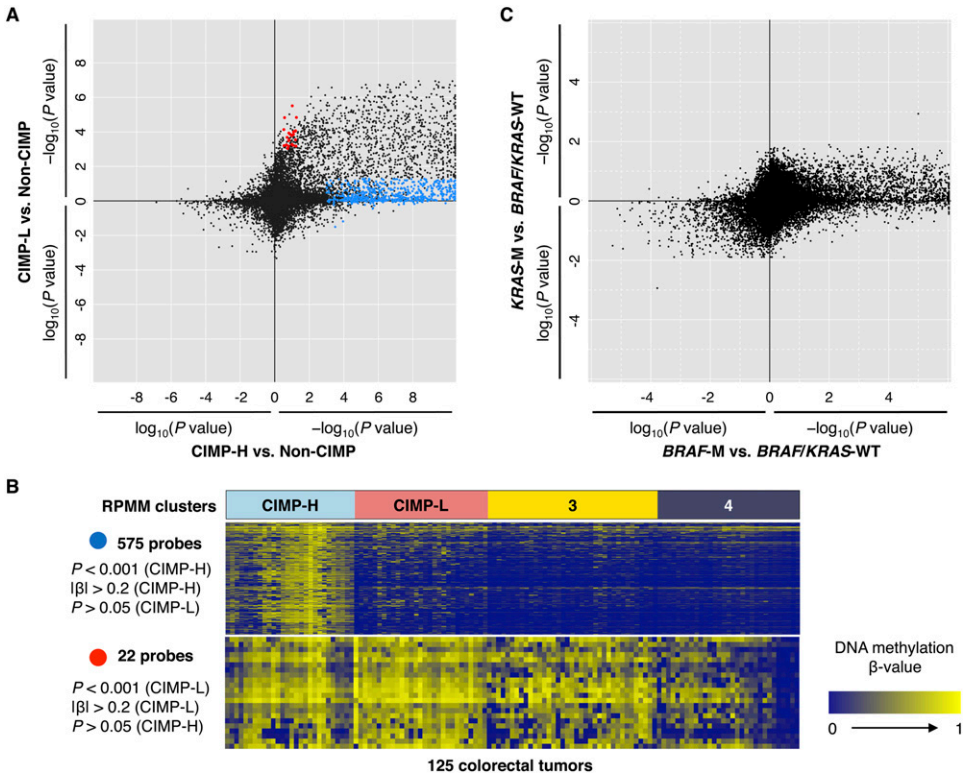

Figure 2. DNA methylation characteristics associated with CIMP-H, CIMP-L, *BRAF*-, and *KRAS* mutant colorectal tumors. (*A*) Comparison of CIMP-H- and CIMP-L-associated DNA methylation profiles. Each data point represents the $\log_{10}$-transformed FDR-adjusted *P*-value comparing DNA methylation in CIMP-H (n = 28) versus non-CIMP tumors (n = 68) (*x*-axis) and in CIMP-L (n = 29) versus non-CIMP tumors (n = 68) (*y*-axis) for each Infinium DNA methylation probe. For the probes with higher mean DNA methylation in CIMP-H or CIMP-L tumors compared to non-CIMP tumors, −1 is multiplied by $\log_{10}$(FDR-adjusted *P*-value), providing positive values. The blue and red points highlight probes that are significantly hypermethylated in CIMP-H and CIMP-L tumors compared to non-CIMP tumors, respectively. (*B*) Heatmap representing Infinium DNA methylation β-values for 575 CpG sites that are significantly hypermethylated in CIMP-H compared with non-CIMP tumors (*top*) and 22 CpG sites that are significantly hypermethylated in CIMP-L compared with non-CIMP tumors (*bottom*). The four DNA methylation-based subgroups are indicated *above* the heatmaps. A color gradient from dark blue to yellow was used to represent the low and high DNA methylation β-values, respectively. (*C*) Comparison of *BRAF* mutant- and *KRAS* mutant-associated DNA hypermethylation signatures in CRC. The $\log_{10}$-transformed FDR-adjusted *P*-value for each probe is plotted for tumors harboring *KRAS* mutations (*KRAS*-M) (n = 34) versus *BRAF*/*KRAS* wild-type (n = 74) (*y*-axis) and those containing *BRAF* mutations (*BRAF*-M) (n = 17) versus *BRAF*/*KRAS* wild-type (n = 74) (*x*-axis). For the probes with higher mean DNA methylation β-values in *BRAF* or *KRAS* mutant tumors compared to wild-type tumors, −1 is multiplied by $\log_{10}$(FDR-adjusted *P*-value), providing positive values.

DNA methylation in a number of CpG sites irrespective of *KRAS* mutation status (Fig. 3C,D). These observations highlight the involvement of more complex molecular mechanisms in driving these DNA methylation clusters.

Sequence characteristics of CIMP-associated gene promoters

We next classified gene promoters that acquired cancer-specific DNA methylation into three categories based on their DNA methylation level profiles across colorectal tumor subtypes (see the Methods section and Supplemental Table 4): (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; and (3) non-CIMP cancer-specific DNA methylation. For comparison, we included 500 gene promoters in two additional groups that did not exhibit cancer-specific DNA methylation profiles and were either constitutively methylated or unmethylated across tumor and adjacent non-tumor colonic tissue samples (Fig. 4).

We explored whether the distinction between these groups of promoters can be attributable to simple structural and sequence characteristics. The majority of genes in all three groups that exhibited cancer-specific DNA methylation as well as the genes that were constitutively unmethylated in normal and tumor tissues are located within CpG islands defined by Takai and Jones (2002) (Fig. 4). We did not observe significant differences in the overall distribution with respect to the CpG observed-to-expected ratio, G:C content, and CpG island length among these four groups of DNA sequences (Supplemental Fig. 5A–C). Therefore, these DNA sequence characteristics do not discriminate among CIMP-associated, non-CIMP-associated, and constitutively unmethylated sequences.

We also considered that specific sequence motifs or repeat sequences surrounding CpG islands may have a role in differential DNA hypermethylation specifically in CIMP tumors. We did not find enrichment or depletion of any di- or tetranucleotide sequences and known transcription factor binding sites in the CIMP-associated CpG islands (data not shown). Recently, Estecio et al. (2010) reported that retrotransposons are more frequently associated with CpG islands that are resistant to DNA hypermethylation than those that are susceptible to DNA hypermethylation. Consistent with their observations, we found that the distances of Infinium DNA methylation probes to the nearest *Alu* repetitive

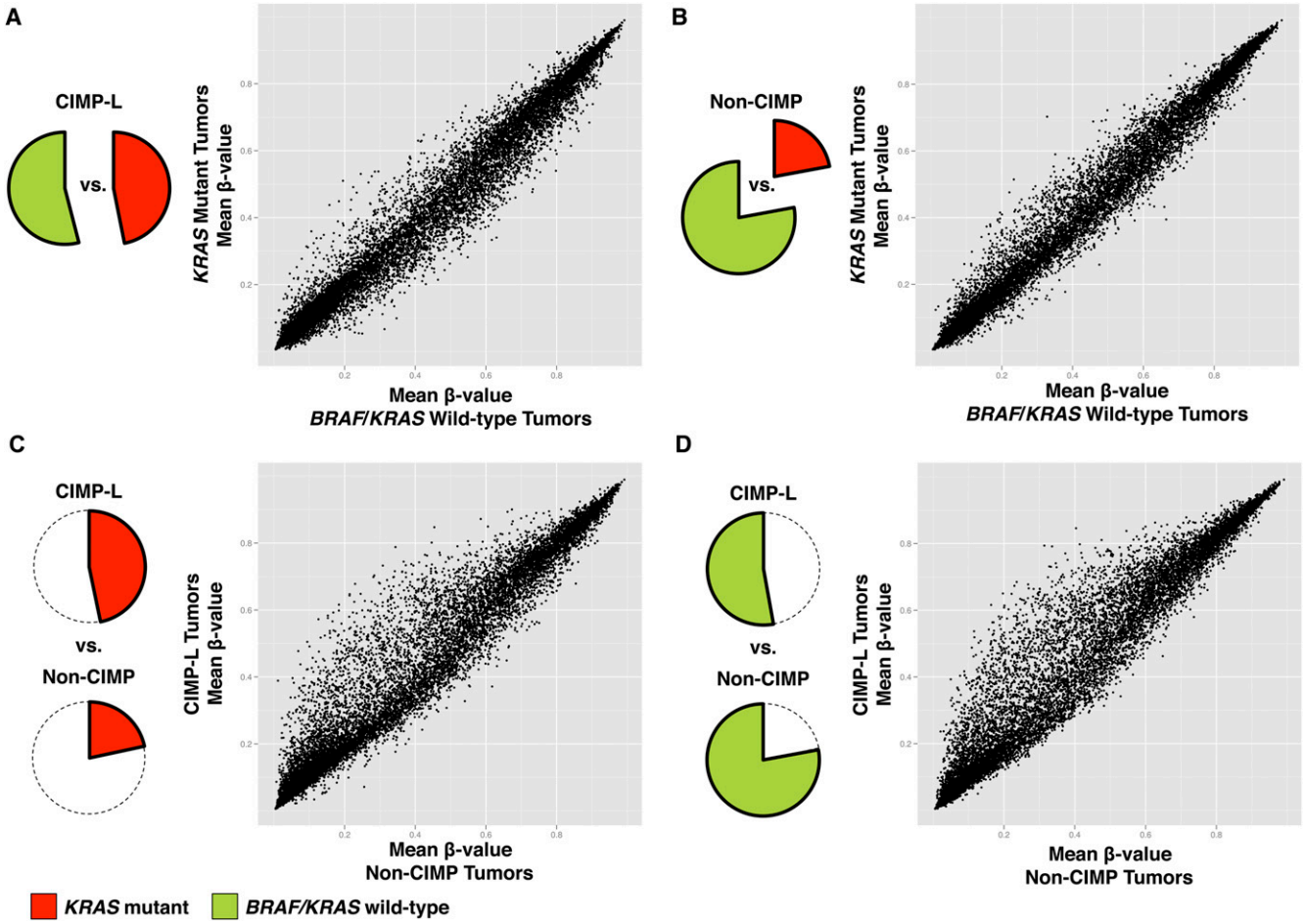

Figure 3. CIMP-L-associated DNA hypermethylation occurs independent of *KRAS* mutation status in CRC. CIMP-L and non-CIMP tumors were subdivided by their *KRAS* and *BRAF* mutation status (*KRAS* mutant or *BRAF*/*KRAS* wild-type), and mean DNA methylation β-values were compared between each group. Scatterplots comparing mean DNA methylation β-values between (*A*) *KRAS* mutant and *BRAF*/*KRAS* wild-type tumors within the CIMP-L subgroup; (*B*) *KRAS* mutant and *BRAF*/*KRAS* wild-type tumors within the non-CIMP subgroups; (C) *KRAS* mutant, CIMP-L tumors versus *KRAS* mutant, non-CIMP tumors; and (*D*) *BRAF*/*KRAS* wild-type, CIMP-L tumors compared with non-CIMP tumors with the same genotype.

element were significantly different between cancer-specifically methylated DNA promoter sequences (median distance: 4300 bp) and those that do not exhibit cancer-specific DNA methylation changes (median distance: 1730 bp; $P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) (Supplemental Fig. 5D). Similarly, we found that cancer-specifically methylated DNA promoter sequences show a greater median distance to LINE repetitive elements compared with those that do not show cancer-specific DNA methylation changes (3880 bp vs. 2710 bp; $P = 1.9 \times 10^{-13}$, Wilcoxon rank-sum test). Interestingly, we observed that differences in the proximity to *Alu* repeat sequences between CIMP-H-associated and non-CIMP-associated promoters are statistically significant with median distances of 3410 bp and 4730 bp, respectively ($P = 1.8 \times 10^{-6}$, Wilcoxon rank-sum test) (Supplemental Fig. 5D). However, we did not find such significant differences for LINE repetitive elements between CIMP-H-associated and non-CIMP-associated promoters ($P = 0.18$).

We next identified the trimethylation status of histone H3 lysine 4 (H3K4me3) and histone H3 lysine 27 (H3K27me3) in human ES cells for genes in the five classification groups described above using a previously published data set (Ku et al. 2008). We found that the genes that are constitutively unmethylated across tumor and adjacent-normal tissue samples are highly enriched for H3K4me3, whereas those that are constitutively methylated are enriched for chromatin states with neither marks in ES cells (Fig. 4). As has previously been reported, the fraction of genes that coincide with ES-cell bivalent domains is substantially higher for the genes that undergo cancer-specific DNA methylation than those that are constitutively methylated or unmethylated across tumor and adjacent non-tumor colonic tissues. We found that >50% of colorectal cancer-specific DNA hypermethylation occurs at ES-cell bivalent domains. However, the proportion of the ES-cell bivalent domains among CIMP-associated and non-CIMP-associated genes is similar, suggesting that the features associated with these targets are not specific for CIMP-positive tumors or CIMP genes, but general features of colorectal cancer (Fig. 4).

Identification of diagnostic CIMP-associated DNA methylation gene marker panels

Next, we developed diagnostic DNA methylation gene marker panels to identify CIMP (CIMP-H and CIMP-L), as well as to segregate CIMP-H tumors from CIMP-L tumors based on the

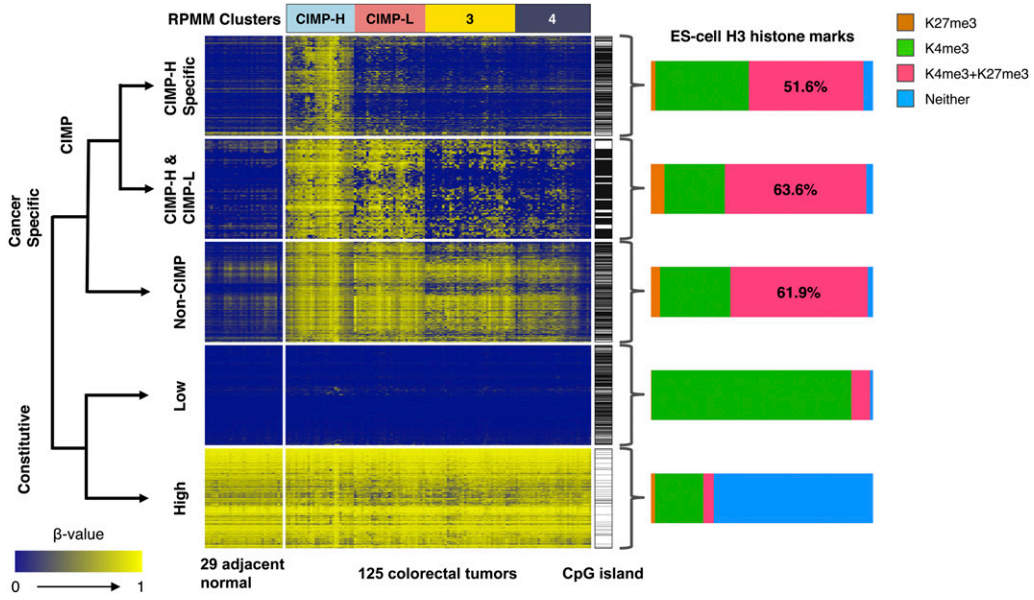

Figure 4. ES-cell histone marks associated with genes in the five classification groups described in the text. Shown are heatmap representations of DNA methylation β-values for unique gene promoters that belong to five different categories: (1) CIMP-H specific: CIMP-associated DNA methylation markers specific for the CIMP-H subgroup only ($n = 415$ genes); (2) CIMP-H & CIMP-L: CIMP-specific DNA methylation shared between the CIMP-H and CIMP-L subgroups ($n = 73$ genes); (3) Non-CIMP: Cancer-specific DNA methylation but outside of the CIMP context ($n = 547$ genes); (4) Constitutive-Low: Constitutively unmethylated genes in both tumor and adjacent normal tissue samples ($n = 500$ genes); (5) Constitutive-High: Constitutively methylated in both tumor and adjacent normal tissue samples ($n = 500$ genes). Genes containing CpG islands defined by Takai and Jones (2002) are indicated by horizontal black bars immediately to the *right* of each heatmap. The bar charts to the *right* of each heatmap show the proportion of gene promoters with occupancy of histone H3 lysine 4 trimethylation (K4me3) and/or histone H3 lysine 27 trimethylation (K27me3) in human ES cells. Probes that do not have this histone mark information (listed in Supplemental Table 4 as "NA") were not included in the bar chart calculations. The probes in each category are ordered according to the unsupervised hierarchal clustering using a correlation distance metric and average linkage method. The RPMM-based cluster assignments are indicated *above* the heatmaps.

Infinium DNA methylation data (Fig. 5). We identified a CIMP-defining marker panel consisting of *B3GAT2*, *FOXL2*, *KCNK13*, *RAB31*, and *SLIT1*. Using the conditions that DNA methylation of three or more markers qualifies a sample as CIMP, this panel identifies CIMP-H and CIMP-L tumors with 100% sensitivity and 95.6% specificity with 2.4% misclassification using a β-value threshold of ≥0.1. The second marker panel of *FAM78A*, *FSTL1*, *KCNC1*, *MYOCD*, and *SLC6A4* specifically identifies CIMP-H tumors with 100% sensitivity and 100% specificity (0% misclassification) using conditions that three or more markers show DNA methylation β-value threshold of ≥0.1. We classify a tumor sample as CIMP-H if both marker panels are positive (three or more markers with DNA methylation for each panel). We classify a tumor sample as CIMP-L if the CIMP-defining marker panel is positive while the CIMP-H specific panel is negative (0–2 genes methylated).

Effects of DNA hypermethylation on gene expression

Promoter CpG island DNA hypermethylation can lead to transcriptional silencing of the associated gene. However, the majority of cancer-specific CpG island hypermethylation may occur in gene promoters that are not normally expressed and therefore may not be involved in tumor initiation or progression (Widschwendter et al. 2007; Gal-Yam et al. 2008). To examine the extent to which cancer-specific DNA hypermethylation affects gene expression in colorectal tumors, we performed an integrated analysis of promoter DNA methylation and gene expression data from six CIMP-H normal adjacent-tumor pairs and 13 pairs of non-CIMP tumors and adjacent non-tumor colonic tissues. We found that 7.3% of genes that showed DNA hypermethylation ($|\Delta\beta| > 0.20$) in CIMP-H tumors also showed more than a twofold reduction in gene

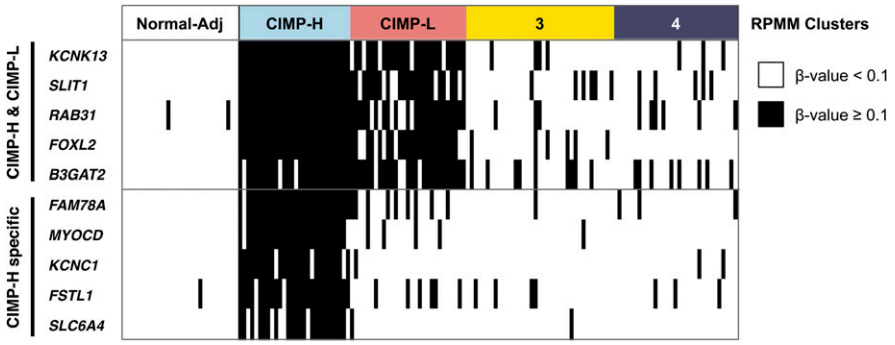

Figure 5. Diagnostic CIMP-defining gene marker panel based on the Infinium DNA methylation data. A dichotomous heatmap of the Infinium DNA methylation data is shown. (Black bars) DNA methylation β-value ≥0.1; (white bars) DNA methylation β-value <0.1. The panel of five markers shown on the *top* (CIMP-H & CIMP-L) is used to identify CIMP-H and CIMP-L tumors. The panel of five markers shown on the *bottom* (CIMP-H specific) is used to specifically identify CIMP-H tumors.

expression (Fig. 6A,B). We identified 464 genes that are down-regulated more than twofold in CIMP-H tumors compared with adjacent normal tissue (Fig. 6A). We found that 112 genes (24%) that are down-regulated in CIMP-H are directly associated with promoter DNA hypermethylation (Supplemental Table 5).

Furthermore, we identified 12 genes that are both down-regulated and cancer-specifically hypermethylated in both CIMP-H and non-CIMP tumors (Fig. 6C; Supplemental Table 5). DNA hypermethylation and transcriptional silencing of these genes may play a critical role in the development of CRC, irrespective of molecular subgroups. These include *SFRP1* and *SFRP2*, which function as negative regulators of Wnt signaling and have been proposed as epigenetic gatekeeper genes in colorectal tumorigenesis (Baylin and Ohm 2006). We validated the DNA methylation and gene expression findings for *SFRP1* and *TMEFF2* using MethyLight and quantitative RT-PCR (qRT-PCR) technologies, respectively (Supplemental Fig. 6).

Intriguingly, we also identified 48/112 genes that are down-regulated in both CIMP-H and non-CIMP tumors compared with the matched adjacent normal colon. However, substantial increases in promoter DNA methylation for these genes were observed only in CIMP-H tumors. We confirmed this finding for the *LMOD1* gene using MethyLight and qRT-PCR technologies (Supplemental Fig. 6). *LMOD1* has been found to be somatically mutated in human cancers and cancer cell lines (http://www.sanger.ac.uk/genetics/CGP/cosmic/). However, DNA hypermethylation of this gene has not yet been reported. These findings indicate that genetic or other epigenetic mechanisms such as chromatin modifications might be involved in silencing of these genes in non-CIMP tumors.

Discussion

CRC can be classified based on various molecular features. Identification and characterization of these subtypes has been not only essential to better understand the disease (Jass 2007), but also valuable in selection of optimal drug treatments, prediction of patient survival, and discovery of risk factors linked to a particular subtype (Walther et al. 2009; Limsui et al. 2010). In this study, we used the Illumina Infinium HM27 DNA methylation assay to investigate DNA methylation-based subgroups in CRC. This BeadArray platform interrogates the gene promoter DNA methylation of all 14,495 consensus coding DNA sequence (CCDS) genes in multiple samples simultaneously and is therefore suitable for a study requiring large-scale promoter DNA methylation profiling of a large number of samples (Bibikova 2009). Using this platform, we identified four DNA methylation subgroups of CRC based on model-based unsupervised cluster analyses. Importantly, the genetic and clinical correlations observed with each subtype suggest that they represent biologically distinct subgroups.

One subgroup designated here as CIMP-H contained all of the CIMP-positive tumors characterized by the MethyLight five-marker panel previously developed in our laboratory (Fig. 1; Weisenberger et al. 2006). Other features associated with the CIMP-H subgroup we described here are also in agreement with those observed in the CIMP1 subtype (Shen et al. 2007) and the high-methylation epigenotype (HME) (Yagi et al. 2010) described previously.

We identified six CIMP-H tumors based on the Infinium DNA methylation data that did not meet the criteria for CIMP using the MethyLight five-gene panel. The MethyLight-based marker panel was developed based on the screening of 195 MethyLight markers (Weisenberger et al. 2006). In our present study, we measured DNA methylation at a much larger number of loci using the Illumina Infinium DNA methylation platform (27,578 CpG sites located at 14,495 gene promoters). The additional loci present on the array probably more accurately identify CIMP tumors, compared to the conventional MethyLight-based five-marker panel. This increased accuracy is likely a reflection of both the inclusion of additional markers that are more tightly associated with CIMP and the mere

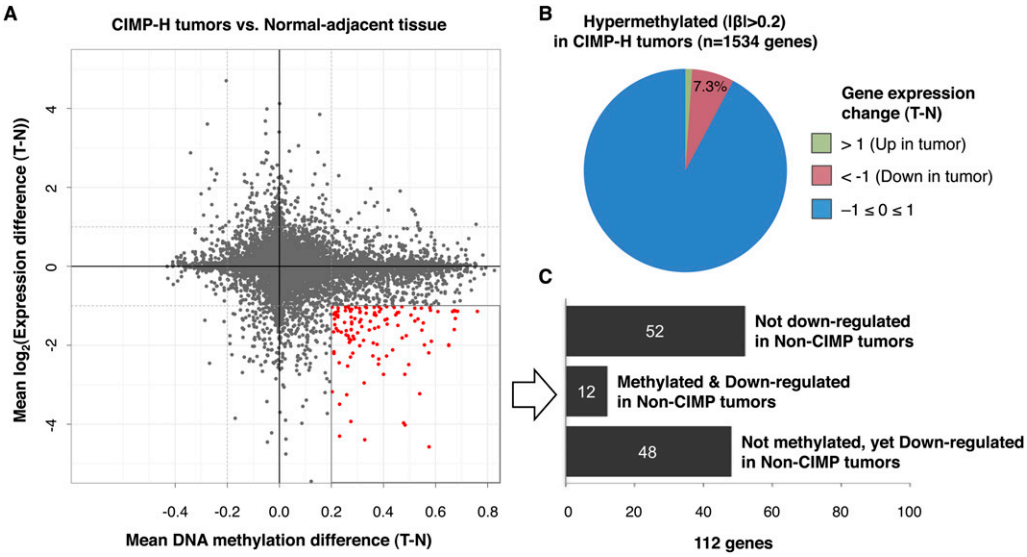

Figure 6. Integrated analysis of gene expression and promoter DNA methylation changes between colorectal tumors and matched normal adjacent tissues. (*A*) Mean DNA methylation β-value differences between CIMP-H tumors and matched normal colonic tissues ($n = 6$) are plotted on the *x*-axis, and mean $\log_2$-transformed gene expression value differences are plotted on the *y*-axis for each gene. Red data points highlight those genes that are hypermethylated with a β-value difference >0.20 and show more than twofold decrease in their gene expression levels in CIMP-H tumors. (*B*) Pie chart showing the gene expression changes of 1534 hypermethylated genes in CIMP-H tumors compared with adjacent normal tissues. (*C*) Bar chart showing the number of genes that exhibit DNA hypermethylation and/or gene expression changes in non-CIMP tumors among the 112 genes that are hypermethylated and down-regulated in CIMP-H tumors.

fact that a larger number of informative loci will usually outperform a small panel of informative loci. The limited MethyLight panel was designed to be compatible with the cost-effective processing of large numbers of formalin-fixed, paraffin-embedded (FFPE) samples. However, any small panel of markers will likely have some misclassification error in identifying a complex molecular profile, regardless of the composition of the panel. Nevertheless, we have found our five-marker CIMP panel to be very useful in large-scale studies of FFPE samples, and thus in this study, we propose new diagnostic DNA methylation marker panels to identify CIMP (CIMP-H and CIMP-L), as well as to segregate CIMP-H tumors from CIMP-L tumors (Fig. 5).

Ogino et al. (2006) proposed the CIMP-low subgroup, which showed DNA hypermethylation of CIMP-defining markers despite a low frequency and enrichment for *KRAS* mutations. Here, we identified the CIMP-L subgroup through a genome-scale approach and provided a comprehensive DNA methylation profile of these tumors. Importantly, the CIMP-L-associated DNA methylation appears to occur only at a subset of CIMP-H-associated sites, as we did not find evidence for strong CIMP-L-specific DNA methylation at a unique set of CpG sites. Moreover, we found that although *KRAS* mutations are enriched in CIMP-L tumors, this subtype may not be driven by *KRAS* mutations, since DNA hypermethylation profiles in *KRAS* wild-type and mutant tumors within CIMP-L tumors were highly correlated across the CpG sites we examined. The independence of *KRAS* mutations from CIMP-L status suggests that a more complex molecular signature exists in driving CIMP-L DNA methylation profiles. Recently, we and others have hypothesized that *BRAF* mutations might be favorably selected in the specific environment that CIMP creates (Hinoue et al. 2009; Suzuki et al. 2010). Similar mechanisms may also result in the enrichment of *KRAS* mutations in the CIMP-L subgroup.

Shen et al. (2007) reported the CIMP2 subset, along with CIMP1 (CIMP-H) and non-CIMP subsets of CRC, using a 28-gene panel. They found a very strong association of CIMP2 with *KRAS* mutations (92%), together with DNA hypermethylation of several CIMP-H-associated loci. The CIMP2 subgroup may be similar to the CIMP-L subgroup we identified in our study. However, we only detected a *KRAS* mutation frequency of ~50% in CIMP-L tumors. The differences in *KRAS* mutation frequencies between CIMP-L and CIMP2 may arise from differences in the CRC patient collections and in the genomic features and technologies used to analyze DNA methylation subgroups of CRC in both studies.

We did not find a statistically significant association of *MGMT* DNA hypermethylation and CIMP-L status. However, Ogino et al. (2007) reported statistical significance in their recent report. The differences between our results and those of Ogino and colleagues may arise from several sources. First, Ogino and colleagues used a different criterion for classifying CIMP-L tumors. Specifically, they classified a tumor sample as CIMP-L if one or two markers from the MethyLight-based CIMP panel showed DNA methylation. Our CIMP-L classification was based on Infinium DNA methylation data, a more rich resource of CIMP-L gene markers. In addition, possible disparities in the CRC sample collections between the studies, such as ethnic population differences, may contribute to CIMP-L classification differences. Finally, there are differences in sample sizes between both studies, which may also contribute to statistical evaluation of CIMP in both collections of CRC tumors.

We also obtained gene expression profiles in pairs of CIMP-H and non-CIMP tumor-normal adjacent tissues to gain insight into the role of CIMP-specific DNA hypermethylation in colorectal tumorigenesis. Aberrant DNA methylation of promoter CpG islands has been established as an important mechanism that inactivates tumor suppressor genes in cancer (Jones and Baylin 2007). However, many cancer-specific CpG island hypermethylation events are also found in promoter regions of genes that are not normally expressed, and these may represent "passenger" events that do not have functional consequences (Widschwendter et al. 2007; Gal-Yam et al. 2008). We examined effects of CIMP-associated DNA hypermethylation on gene expression. We found that only 7.3% of the CIMP-H-specific DNA methylation markers showed a strong inverse relationship with their gene expression levels. Similar observations have been made in the glioma-CpG island methylator phenotype (G-CIMP) (Noushmehr et al. 2010). Although a larger sample size is required for better estimates, our observations might reinforce the hypothesis that CIMP represents a broad epigenetic control defect that accompanies a large number of "passenger" DNA hypermethylation events (Weisenberger et al. 2006).

We identified 112 genes that showed both promoter DNA hypermethylation and reduction in gene expression in CIMP-H tumors. Importantly, we found that 12 of these genes also showed DNA hypermethylation with a concomitant reduction in gene expression level in non-CIMP tumors, which indicates that aberrant DNA methylation and transcriptional silencing of these genes may be important in the development of CRC, irrespective of molecular subtype. Intriguingly, these include *SFRP1* and *SFRP2*, which function as negative regulators of Wnt signaling. DNA hypermethylation of *SFRP* genes has been observed in the majority of ACFs and colorectal tumors, and these genes have been described as epigenetic gatekeepers in colorectal tumorigenesis (Baylin and Ohm 2006). DNA hypermethylation and transcriptional silencing of other genes such as *TMEFF2* and *SLIT3* have also been reported (Young et al. 2001; Dickinson et al. 2004). However, the functional significance of the inactivation of these genes has not been established in CRC.

Interestingly, we also noticed that of the 112 genes that exhibited DNA hypermethylation and reduced gene expression in CIMP-H tumors, 48 were also silenced in non-CIMP tumors, but without substantial increases in DNA methylation. CIMP status in CRC has been found to be inversely correlated with the occurrence of chromosomal instability (CIN), which is characterized by aneuploidy, gain and loss of subchromosomal genomic regions, and high frequencies of loss of heterozygosity (LOH) (Goel et al. 2007; Cheng et al. 2008). Recently, Chan et al. (2008) identified genes that are inactivated by both genetic mechanisms (mutation or deletion) and DNA hypermethylation in breast and colorectal cancer. They observed that these genetic and epigenetic changes are generally mutually exclusive in a given tumor, and that silencing of these genes was associated with poor clinical outcome (Chan et al. 2008). Together, these genes may act as key tumor suppressor genes in CRC, and the gene-silencing mechanisms can be determined by the underlying molecular pathways involved in colorectal tumorigenesis.

The molecular mechanisms that account for CIMP have not been identified. It has been proposed that CIMP arises through a distinct pathway originating in a variant of hyperplastic polyps and sessile serrated adenomas due to the similar histological and molecular features shared by the CIMP tumors and these lesions (O'Brien 2007). Some individuals and families with hyperplastic polyposis syndrome have an increased risk of developing CIMP CRC, indicating the existence of a genetic predisposition that could lead to CIMP (Young et al. 2007). Environmental exposures

might also influence the risk of developing CIMP CRC. Cigarette smoking was found to be associated with increased risk of developing CIMP CRC in a recent report (Limsui et al. 2010).

Here, we were not able to find characteristic sequence signatures in the CIMP-associated CpG islands. Future studies will be directed toward identifying and characterizing the genomic localization of other chromatin marks or proteins that are involved in organizing higher-order chromatin architecture. Integrated analyses with this information may provide insights into the molecular mechanism of CIMP.

Together, the findings described in our study provide the most comprehensive genome-scale analysis of DNA methylation-based subgroups of CRC to date. The unique DNA methylation profiles in CRC, together with genomic changes, provide a detailed molecular landscape of colorectal tumors. Our findings here have clinical implications on colorectal cancer diagnosis and may be helpful in directing treatment for CRC patients.

Methods

Primary colorectal tissue sample collection and processing

Twenty-five paired colorectal tumor and histologically normal adjacent colonic tissue samples were obtained from colorectal cancer patients who underwent surgical resection at the department of surgery in the Groene Hart Hospital, Gouda, The Netherlands. Tissue samples were stored at −80°C within 1 h after resection. Tissue sections from the surgical resection margin were examined by a pathologist (C.M. van Dijk) by microscopic observation. All patients provided written informed consent for the collection of samples and subsequent analysis. The study was approved by the Institutional Review Board of the Groene Hart Hospital in Gouda, the Leiden University Medical Center, and the University of Southern California. An additional collection of 100 fresh-frozen colorectal tumor samples and four matched histologically normal colonic mucosa tissue samples adjacent to the tumors were obtained from the Ontario Tumor Bank Network (The Ontario Institute for Cancer Research, Ontario, Canada). The tissue collection and analyses were approved by the University of Southern California Institutional Review Board. Genomic DNA and total RNA were extracted simultaneously from the same tissue sample using the TRIzol Reagent (Invitrogen) according to the manufacturer's protocol.

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. Mutations in *TP53* exons 4 through 8 were determined by direct sequencing of PCR products. Samples containing missense mutations, nonsense mutations, splice-site mutations, frameshift mutations, and in-frame deletions were considered positive for a mutation. Additional details including primer sequences are provided in the Supplemental Material.

DNA methylation assay

Details regarding the MethyLight assay are provided in the Supplemental Material. The Illumina Infinium HumanMethylation27 DNA methylation assay technology has been described previously (Bibikova 2009). Briefly, genomic DNA was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research) according to the manufacturer's instructions. We assessed the amount of bisulfite-converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described (Campan et al. 2009). All of the samples in this study passed our QC tests and entered into the Infinium DNA methylation assay pipeline. The Infinium DNA methylation assay was performed at the USC Epigenome Center according to the manufacturer's specifications (Illumina). The Illumina Infinium DNA methylation assay examines the DNA methylation status of 27,578 CpG sites located at promoter regions of 14,495 protein-coding genes and 110 microRNAs. A measure of the level of DNA methylation at each CpG site is scored as beta (β) values ranging from 0 to 1, with values close to 0 indicating low levels of DNA methylation and values close to 1 indicating high levels of DNA methylation (Bibikova 2009). The detection *P*-values measure the difference of the signal intensities at the interrogated CpG site compared with those from a set of 16 negative control probes embedded in the assay. We identified all data points with a detection *P*-value >0.05 as not statistically significantly different from background measurements, and therefore not trustworthy measures of DNA methylation. These data points were replaced by "NA" values as previously described (Noushmehr et al. 2010). The assay probe sequences and detailed information on each interrogated CpG site and the associated genomic characteristics on the HumanMethylation27 BeadChip can be obtained at http://www.illumina.com. All Infinium DNA methylation data are available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE25062.

Gene expression assay

Gene expression assays were performed on 25 pairs of colorectal tumor and adjacent non-tumor colonic tissues using the Illumina Ref-8 whole-genome expression BeadChip (HumanRef-8 v3.0, 24,526 transcripts; Illumina). Scanned image and bead-level data processing were performed using the BeadStudio 3.0.1 software (Illumina). The summarized data for each bead type were then processed using the lumi package in Bioconductor (Du et al. 2008). The data were $\log_2$-transformed and normalized using Robust Spline Normalization (RSN) as implemented in the lumi package. The summarized probe profile data and processed expression data are available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE25070. Additional assay details are provided in the Supplemental Material.

Data filtering and normalization

For the Illumina Infinium DNA methylation data analysis, we masked data points as "NA" for probes that might be unreliable (see the Supplemental Methods). We also identified all data points with a detection *P*-value >0.05 and replaced those with "NA" values. Finally, we excluded probes that are designed for sequences on either the X or Y chromosome. We analyzed the DNA methylation data set that does not contain any "NA"-masked data points. DNA methylation β-values were normalized to eliminate the batch effects. Briefly, the batch means of β-values were brought closer to the overall mean while retaining the original range of DNA methylation data (0 to 1). We used only the tumor samples to calculate the batch means and overall mean in estimating the scaling factor for each batch. For the gene expression analysis, unreliable probes (9%) as described by Barbosa-Morais et al. (2010) were removed from the subsequent analysis.

Unsupervised clustering

We used the recursively partitioned mixture model (RPMM) for the identification of colorectal tumor subgroups based on the Illumina

Infinium DNA methylation data. RPMM is a model-based unsupervised clustering approach developed for beta-distributed DNA methylation measurements that lie between 0 and 1 and implemented as the RPMM Bioconductor package (Houseman et al. 2008). We identified probes that do not contain any "NA"-masked data points and then performed RPMM clustering on 2758 probes (10% of original probes) that showed the most variable DNA methylation levels across the colorectal tumor panel. A fanny algorithm (a fuzzy clustering algorithm) was used for initialization and level-weighted version of Bayesian information criterion (BIC) as a split criterion for an existing cluster as implemented in the R-based RPMM package.

Statistical analysis and data visualization

Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org). The Wilcoxon rank-sum test and the Wilcoxon signed-rank test were used to evaluate the difference in DNA methylation β-values for each probe between two independent groups and between tumor and matched adjacent-normal tissues, respectively. False-discovery rate (FDR)–adjusted *P*-values for multiple comparisons were calculated using the Benjamini and Hochberg approach. The Illumina Infinium DNA methylation β-values were represented graphically using a heatmap, generated by the R/Bioconductor packages gplots and Heatplus. Ordering of the samples within an RPMM class in the heatmaps was obtained by using the function "seriate" in the seriation package. We have created a Sweave document containing the details of our analyses in the Supplemental Material.

Classification and selection of cancer-specific DNA methylation markers

We categorized gene promoters that exhibited cancer-specific DNA methylation into three groups. We selected 415 unique gene promoters that showed significant CIMP-H-specific DNA hypermethylation (FDR-adjusted $P < 0.0001$ for CIMP-H vs. non-CIMP tumors and $P > 0.05$ for CIMP-L vs. non-CIMP tumors) and 73 gene promoters that showed DNA hypermethylation in both CIMP-H and CIMP-L tumors (FDR-adjusted $P < 0.0001$ for CIMP-H vs. non-CIMP and CIMP-L vs. non-CIMP). For the third category, we identified 547 genes that acquired cancer-specific DNA hypermethylation irrespective of CIMP status (FDR-adjusted $P < 0.00001$ for 29 paired tumor vs. adjacent non-tumor tissue) (for a list of genes, see Supplemental Table 4).

Identification of diagnostic CIMP-associated DNA methylation gene marker panels

We first selected the top 20 Infinium DNA methylation probes that are significantly hypermethylated in CIMP (CIMP-H and CIMP-L) compared with non-CIMP tumors based on the Wilcoxon rank-sum test. Using the conditions that a DNA methylation β-value ≥ 0.1 of three or more markers qualifies a sample as CIMP, we determined a five-probe panel that best classifies CIMP (CIMP-H and CIMP-L) by calculating the sensitivity, specificity, and overall misclassification rate for each random combination of the top 20 probes. For the CIMP-H-specific marker panel, we first selected the top 20 probes that are significantly hypermethylated in CIMP-H compared with CIMP-L tumors. We then chose a five-marker panel that showed the best sensitivity, specificity, and overall misclassification rate to classify CIMP-H using the conditions that three or more markers show a DNA methylation β-value threshold of ≥ 0.1.

Integrated analyses of the Illumina Infinium DNA methylation and gene expression data

We selected one probe for each gene that showed the highest absolute mean β-value difference between tumor and adjacent non-tumor colonic tissues. We then merged the DNA methylation and gene expression data set using Entrez Gene IDs using the R "merge" function. We considered expression data points with a detection *P*-value >0.01, computed by BeadStudio software, as not distinguishable from the negative control measurements and therefore not expressed. We used a mean β-value difference ($|\Delta\beta|$) of 0.20 as a threshold for differential DNA methylation. This threshold of $|\Delta\beta| = 0.20$ was determined previously as a stringent estimate of $\Delta\beta$ detection sensitivity across the range of β-values (Bibikova 2009).

Data access

The data discussed in this manuscript have been deposited in the NCBI Gene Expression Omnibus (GEO) and are accessible through GEO Series accession numbers GSE25062 and GSE25070.

Acknowledgments

The work described in this manuscript was supported by U.S. National Institutes of Health grant R01 CA075090 awarded to P.W.L. We acknowledge generous support of colon cancer research from William and Darlene Christy in memory of Stephanie Anne Christy.

References

Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, Tavare S. 2010. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* **38:** e17. doi: 10.1093/nar/gkp942.

Baylin SB, Ohm JE. 2006. Epigenetic gene silencing in cancer—a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer* **6:** 107–116.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. *Cell* **128:** 669–681.

Bibikova M. 2009. Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics* **1:** 177–200.

Campan M, Weisenberger DJ, Trinh B, Laird PW. 2009. MethyLight. *Methods Mol Biol* **507:** 325–337.

Chan AO, Broaddus RR, Houlihan PS, Issa JP, Hamilton SR, Rashid A. 2002. CpG island methylation in aberrant crypt foci of the colorectum. *Am J Pathol* **160:** 1823–1830.

Chan TA, Glockner S, Yi JM, Chen W, Van Neste L, Cope L, Herman JG, Velculescu V, Schuebel KE, Ahuja N, et al. 2008. Convergence of mutation and epigenetic alterations identifies common genes in cancer that predict for poor prognosis. *PLoS Med* **5:** e114. doi: 10.1371/journal.pmed.0050114.

Cheng YW, Pincas H, Bacolod MD, Schemmann G, Giardina SF, Huang J, Barral S, Idrees K, Khan SA, Zeng Z, et al. 2008. CpG island methylator phenotype associates with low-degree chromosomal abnormalities in colorectal cancer. *Clin Cancer Res* **14:** 6005–6013.

Christensen BC, Houseman EA, Godleski JJ, Marsit CJ, Longacker JL, Roelofs CR, Karagas MR, Wrensch MR, Yeh RF, Nelson HH, et al. 2009a. Epigenetic profiles distinguish pleural mesothelioma from normal pleura and predict lung asbestos burden and clinical outcome. *Cancer Res* **69:** 227–234.

Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, et al. 2009b. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* **5:** e1000602. doi: 10.1371/journal.pgen.1000602.

Christensen BC, Kelsey KT, Zheng S, Houseman EA, Marsit CJ, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Kushi LH, et al. 2010. Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake. *PLoS Genet* **6:** e1001043. doi: 10.1371/journal.pgen.1001043.

Christensen BC, Smith AA, Zheng S, Koestler DC, Houseman EA, Marsit CJ, Wiemels JL, Nelson HH, Karagas MR, Wrensch MR, et al. 2011. DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. *J Natl Cancer Inst* **103:** 143–153.
Dickinson RE, Dallol A, Bieche I, Krex D, Morton D, Maher ER, Latif F. 2004. Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers. *Br J Cancer* **91:** 2071–2078.
Du P, Kibbe WA, Lin SM. 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24:** 1547–1548.
Estecio MR, Gharibyan V, Shen L, Ibrahim AE, Doshi K, He R, Jelinek J, Yang AS, Yan PS, Huang TH, et al. 2007. LINE-1 hypomethylation in cancer is highly variable and inversely correlated with microsatellite instability. *PLoS ONE* **2:** e399. doi: 10.1371/journal.pone.0000399.
Estecio MR, Gallegos J, Vallot C, Castoro RJ, Chung W, Maegawa S, Oki Y, Kondo Y, Jelinek J, Shen L, et al. 2010. Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer. *Genome Res* **20:** 1369–1382.
Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301:** 89–92.
Gal-Yam EN, Egger G, Iniguez L, Holster H, Einarsson S, Zhang X, Lin JC, Liang G, Jones PA, Tanay A. 2008. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci* **105:** 12979–12984.
Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, Ehrlich M. 1983. The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res* **11:** 6883–6894.
Goel A, Nagasaka T, Arnold CN, Inoue T, Hamilton C, Niedzwiecki D, Compton C, Mayer RJ, Goldberg R, Bertagnolli MM, et al. 2007. The CpG island methylator phenotype and chromosomal instability are inversely correlated in sporadic colorectal cancer. *Gastroenterology* **132:** 127–138.
Hinoue T, Weisenberger DJ, Pan F, Campan M, Kim M, Young J, Whitehall VL, Leggett BA, Laird PW. 2009. Analysis of the association between CIMP and $BRAF^{V600E}$ in colorectal cancer by DNA methylation profiling. *PLoS ONE* **4:** e8357. doi: 10.1371/journal.pone.0008357.
Houseman EA, Christensen BC, Yeh RF, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, et al. 2008. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9:** 365. doi: 10.1186/1471-2105-9-365.
Ibrahim AE, Arends MJ, Silva AL, Wyllie AH, Greger L, Ito Y, Vowler SL, Huang TH, Tavare S, Murrell A, et al. 2011. Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. *Gut* **60:** 499–508.
Jass JR. 2007. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **50:** 113–130.
Jiang X, Tan J, Li J, Kivimae S, Yang X, Zhuang L, Lee PL, Chan MT, Stanton LW, Liu ET, et al. 2008. DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell* **13:** 529–541.
Jones PA, Baylin SB. 2007. The epigenomics of cancer. *Cell* **128:** 683–692.
Kondo Y, Shen L, Cheng AS, Ahmed S, Boumber Y, Charo C, Yamochi T, Urano T, Furukawa K, Kwabi-Addo B, et al. 2008. Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet* **40:** 741–750.
Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4:** e1000242. doi: 10.1371/journal.pgen.1000242.
Leggett B, Whitehall V. 2010. Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology* **138:** 2088–2100.
Limsui D, Vierkant RA, Tillmans LS, Wang AH, Weisenberger DJ, Laird PW, Lynch CF, Anderson KE, French AJ, Haile RW, et al. 2010. Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *J Natl Cancer Inst* **102:** 1012–1022.
Marsit CJ, Christensen BC, Houseman EA, Karagas MR, Wrensch MR, Yeh RF, Nelson HH, Wiemels JL, Zheng S, Posner MR, et al. 2009. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. *Carcinogenesis* **30:** 416–422.
Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA, Kelsey KT. 2011. DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. *J Clin Oncol* **29:** 1133–1139.
Miranda TB, Jones PA. 2007. DNA methylation: The nuts and bolts of repression. *J Cell Physiol* **213:** 384–390.
Monti S, Tamayo P, Mesirov J, Golub T. 2003. Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning J* **52:** 91–118.
Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, et al. 2010. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17:** 510–522.
O'Brien MJ. 2007. Hyperplastic and serrated polyps of the colorectum. *Gastroenterol Clin North Am* **36:** 947–968.
Ogino S, Kawasaki T, Kirkner GJ, Loda M, Fuchs CS. 2006. CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: Possible associations with male sex and *KRAS* mutations. *J Mol Diagn* **8:** 582–588.
Ogino S, Kawasaki T, Kirkner GJ, Suemoto Y, Meyerhardt JA, Fuchs CS. 2007. Molecular correlates with MGMT promoter methylation and silencing support CpG island methylator phenotype-low (CIMP-low) in colorectal cancer. *Gut* **56:** 1564–1571.
Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, et al. 2007. A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* **39:** 237–242.
Pino MS, Chung DC. 2010. The chromosomal instability pathway in colon cancer. *Gastroenterology* **138:** 2059–2072.
Rodriguez J, Munoz M, Vives L, Frangou CG, Groudine M, Peinado MA. 2008. Bivalent domains enforce transcriptional memory of DNA methylated genes in cancer cells. *Proc Natl Acad Sci* **105:** 19809–19814.
Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, et al. 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39:** 232–236.
Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, Hernandez NS, Chen X, Ahmed S, Konishi K, et al. 2007. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci* **104:** 18654–18659.
Suzuki H, Igarashi S, Nojima M, Maruyama R, Yamamoto E, Kai M, Akashi H, Watanabe Y, Yamamoto H, Sasaki Y, et al. 2010. IGFBP7 is a p53-responsive gene specifically silenced in colorectal cancer with CpG island methylator phenotype. *Carcinogenesis* **31:** 342–349.
Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci* **99:** 3740–3745.
Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D. 2009. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer* **9:** 489–499.
Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D, et al. 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* **38:** 787–793.
Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan M, Young J, Jacobs I, et al. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* **39:** 157–158.
Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113.
Yagi K, Akagi K, Hayashi H, Nagae G, Tsuji S, Isagawa T, Midorikawa Y, Nishimura Y, Sakamoto H, Seto Y, et al. 2010. Three DNA methylation epigenotypes in human colorectal cancer. *Clin Cancer Res* **16:** 21–33.
Young J, Jass JR. 2006. The case for a genetic predisposition to serrated neoplasia in the colorectum: hypothesis and review of the literature. *Cancer Epidemiol Biomarkers Prev* **15:** 1778–1784.
Young J, Biden KG, Simms LA, Huggard P, Karamatic R, Eyre HJ, Sutherland GR, Herath N, Barker M, Anderson GJ, et al. 2001. HPP1: A transmembrane protein-encoding gene commonly methylated in colorectal polyps and cancers. *Proc Natl Acad Sci* **98:** 265–270.
Young J, Jenkins M, Parry S, Young B, Nancarrow D, English D, Giles G, Jass J. 2007. Serrated pathway colorectal cancer in the population: genetic consideration. *Gut* **56:** 1453–1459.

Received November 8, 2010; accepted in revised form May 6, 2011.