Supplemental Methods

COHCAP – 1 Group Workflows

Average by CpG Site (recommended): Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. Beta values are averaged among all samples within a group for a given CpG site, and methylated and unmethylated thresholds are used to identify CpG sites that are methylated or unmethylated, and ambiguous CpG sites are removed. CpG sites with ambiguous average beta values are filtered out. CpG islands (as defined by the annotation file, which also defines the mapping between CpG islands and genes) are considered for statistical analysis if they possess a minimum number of filtered CpG sites (default = 4). CpG island p-values are calculated by Fisher's Exact test comparing methylated and unmethylated CpG site counts within a CpG island genome-wide methylated and unmethylated CpG site counts. False discovery rate (FDR) values are then calculated using the method of Benjamini and Hochberg [45]. Integration with gene expression data is not possible with this workflow.

Average by CpG Island: Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. Beta values are averaged among all CpG sites within a group, and methylated and unmethylated thresholds are used to identify CpG sites that are methylated or unmethylated, and ambiguous CpG sites are removed. CpG sites with ambiguous average beta values are filtered out. Finally, beta values are averaged among CpG sites within a CpG islands. DNA methylation and gene expression data (represented by a table of expression / intensity values) are integrated by calculating correlations for paired samples.

COHCAP – 2 Group Workflows

Average by CpG Site: Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. If samples are paired, then CpG site p-values are calculated via ANOVA F-statistic. Otherwise, CpG site p-values are calculated via t-test. False discovery rate (FDR) values then calculated using the method of Benjamini and Hochberg [45]. Beta values are then averaged among all samples within a group for a given CpG site. CpG sites are then filtered based upon average beta values (above methylation cutoff in one group and below methylation cutoff in the other group), p-value, and FDR. CpG islands (as defined by the annotation file, which also defines the mapping between CpG islands and genes) are considered for statistical analysis if they possess a minimum number of filtered CpG sites (default = 4). CpG site p-values are calculated via ANOVA F-statistic for averaged beta values (using a 2-way ANOVA considering the influence of group and CpG site pairs on beta value variation). FDR values then calculated using the method of Benjamini and Hochberg [45]. DNA methylation and gene expression data (represented by a table containing fold-change, p-value, and FDR values) are integrated by filtering for islands / genes meeting various criteria (e.g. methylated threshold, unmethylated threshold, CpG island p-value, CpG island FDR, expression fold-change, expression p-value, and/or expression FDR), and then listing genes with an inverse relationship between change in DNA methylation and change in gene expression).

Average by CpG Island (recommended): Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. If samples are paired, then CpG site p-values are calculated via ANOVA F-statistic. Otherwise, CpG site p-values are calculated via t-test. False discovery rate (FDR) values then calculated using the method of Benjamini and Hochberg [45]. CpG sites are then filtered based upon average beta values (above methylation cutoff in one group and below methylation cutoff in the other group), p-value, and FDR. If a CpG islands (as defined by the annotation file, which also defines the mapping between CpG islands and genes) contains a minimum number of filtered CpG sites

(default = 4), then average beta values are calculated between filtered CpG sites in each CpG island (per sample). CpG island beta values are then treated like CpG site beta values for statistical analysis. DNA methylation and gene expression data (represented by a table of expression / intensity values) are integrated by calculating correlations for paired samples.

COHCAP – 3+ Group Workflows

Average by CpG Site: Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. CpG site p-values are calculated via ANOVA F-statistic (if samples are paired, 2-way ANOVA is used; otherwise, 1-way ANOVA is used). False discovery rate (FDR) values then calculated using the method of Benjamini and Hochberg [45]. CpG sites are then filtered based upon p-value and FDR, and beta values are then averaged among all samples within a group for a given CpG site. CpG islands (as defined by the annotation file, which also defines the mapping between CpG islands and genes) are considered for statistical analysis if they possess a minimum number of filtered CpG sites (default = 4). CpG site p-values are calculated via ANOVA F-statistic for averaged beta values (using a 2-way ANOVA considering the influence of group and CpG site pairs on beta value variation). FDR values then calculated using the method of Benjamini and Hochberg [45]. Integration with gene expression data is not possible with this workflow.

Average by CpG Island (recommended): Table of beta values is annotated using annotation file and sample description file. Values with detection p-value greater than cutoff are censored. Histogram, dendrogram, and principal component analysis (PCA) plot are created based upon beta values for samples listed in sample description file. CpG site p-values are calculated via ANOVA F-statistic (if samples are paired, 2-way ANOVA is used; otherwise, 1-way ANOVA is used). False discovery rate (FDR) values then calculated using the method of Benjamini and Hochberg [45]. CpG sites are then filtered based upon p-value and FDR. If a CpG islands (as defined by the annotation file, which also defines the mapping between CpG islands and genes) contains a minimum number of filtered CpG sites

(default = 4), then average beta values are calculated between filtered CpG sites in each CpG island (per sample). CpG island beta values are then treated like CpG site beta values for statistical analysis. DNA methylation and gene expression data (represented by a table of expression / intensity values) are integrated by calculating correlations for paired samples.

Sample Processing

When raw Illumina 450k methylation data was available, the data was normalized in Illumina® Genome Studio™ (V2011.1) using a background correction normalized to Illumina controls (without providing any grouping information – each sample was identified as a separate group for normalization purposes). Processed beta values for publicly available data [10] were used for COHCAP analysis..

Processed RPKM (Reads Per Kilobase per Million) expression levels were used for the TCGA RNA-Seq data. Sample TCGA-BH-A0AW-01A was removed from the ER+ vs. ER- analysis because it showed a high proportion of probes with detection p-value < 0.05. This sample was not present in the paired cancer vs. normal analysis. All other samples had >99% of probes with a detection p-value < 0.05. This processed RNA-Seq data was used in COHCAP for the "Average by Island" workflow. Partek® Genomics Suite™ (Version 6.6; Partek, Inc., St. Louis, MO) was used for calculating fold-change, p-value, and FDR values for the "Average by Site" COHCAP workflow (for the TCGA data as well as the HCT116 data published with this study). The Human Gene 1.0 ST Array used to measure parental and mutant gene expression in this study was processed via RMA normalization [56]. When performing integration via overlap, a gene was defined as differentially expressed if it showed at least at 1.5 fold-change and an FDR less than 0.05. Fold-change values are calculated based upon the least-squares mean, p-values were calculated using 1-way ANOVA with appropriate linear contrast, and FDR values were calculated using the method of Benjamini and Hochberg [45].

Raw BS-Seq data [36] was aligned using Bismark [43], and sites with at least 10x coverage were included in the .bed files used in for COHCAP analysis.

The novel methylation and gene expression data presented in this study can be downloaded from GEO (GEO SuperSeries GSE42310; expression in GSE42307, HCT116 methylation in GSE42308, HES-2 450k methylation in GSE42707, and HES-2 MIRA methylation in GSE42734).

HCT116 Dataset and Validation of Differential Methylation

Gene expression analysis was performed using RNA extracted with the RNeasy kit (Qiagen) from HCT116 human colon cancer cell lines. The microarray sample preparation was carried out using Ambion's WT Expression kit (Life Technologies, Carlsbad, CA) and Affymetrix's GeneChip Terminal labeling system at the COH Functional Genomics Core using a procedure described previously [57]. Briefly, 100 ng of total RNA was used to start the first strand cDNA synthesis using random primers plus polyT7 promoter. The RNA quality was checked using Agilent Bioanalzer nano 6000 and all RNA had a RIN > 9 for this study before the 1st stand synthesis. After the 2nd strand cDNA synthesis, the antisense cRNA was carried using T7 RNA polymerase. Then, 10 µg of cRNA was used to start the 2nd cycle of cDNA synthesis using random primers plus dUTP and dNTP mix. The single-strand cDNA was fragmented and then end-labeled with biotinylated nucleotides in the presence of terminal deoxynucleotidyl transferase (TdT) using Affymetrix WT Terminal Labeling kit. Five µg of labeled single stranded cDNA was hybridized with Affymetrix Human Gene 1.0 ST arrays the arrays were scanned using Affymetrix GeneChip Scanner 3000 7G.

Genomic DNA was prepared from HCT116 human colon cancer cell lines using the DNeasy® Tissue kit following manufacturer's instructions (Qiagen). DNA was extracted from HCT116 human colon cancer cell lines using Qiagen kit. Illumina MethylHuman450k beadchip methylation analysis was performed following Illumina Infinium HD methylation assay protocol at the COH Functional Genomics Core. Briefly, 500 ng high quality genomic DNA of each sample was treated with Zymo EZ DNA methylation kit (Zymo Research) to convert unmethylated cytosines to uracil, while leaving methylated cytosines unchanged for methylation analysis. The bisulfate-converted DNA was denatured with 0.1 N NaOH and amplified at 37°C for 20-24 hours. After fragmentation, the amplified DNA was precipitated with 2-propanol. The DNA was resuspended in 46 µl RA1 buffer and 15 µl resuspended DNA was loaded

into each beadchip section. The beadchip was assembled into the Hyb chamber and incubated in a 48°C degree in Illumina hybridization oven for at least 16 hours. After washing, the beadchip was assembled into a flow-through chamber and single base extension and staining was performed at 44°C on the chamber rack. After washing, the stained beadchip was coated with XC4 buffer and scanned using Illumina's HiScanSQ scanner.

Validation of gene expression levels was determined by real-time RT-PCR. Briefly, total RNA was reverse transcribed to cDNA using iScript cDNA Synthesis Kit (Bio-Rad), and real-time RT-PCR reactions were performed using iQ SYBR Green supermix (Bio-Rad) on a DNA Engine thermal cycler equipped with Chromo4 detector (Bio-Rad). Gene specific primer sets were purchased from RealTimePrimers.

EpiTect® MethylPCR assay (Qiagen) was performed to validate DNA methylation levels. Unmethylated or methylated DNA was selectively digested by methylation-sensitive (cuts unmethylated and partially methylated DNA, leaving only hypomethylated DNA) and methylation-dependent restriction enzymes (cuts any methylated DNA, leaving only unmethylated DNA) according to the manufacturer's instructions. The remaining DNA after digestion was quantified by real-time PCR using primers detecting methylation status of the CpG islands associated with the individual gene. The relative concentrations of differentially methylated DNA were determined by comparing the amount of each digest with that of mock digest (no enzyme added), using the software provided by the manufacturer (Qiagen). In all cases, validation was performed in triplicate.

MIRA Analysis Comparison

HES-2 cell line DNA was prepared as described in Tompkins et al. [58], and the microarray sample was processed at the UCLA DNA Microarray Core. The data for the MIRA sample (GSE42734) was processed using NimbleScan (v 2.6.0.0) to compute ratio files and find peaks (sliding window width = 750 bp, mimimum $-\log_{10}$ p-value cutoff = 2.0, maximum spacing between nearby probes within peak = 500 bp, minimum probes per peak = 2). The NimbleGen MeDIP array (Human DNA Methylation 3 x 720K CpG Island Plus RefSeq Promoter Array) used for the MIRA assay defined genome coordinates based

upon hg18 (while the UCSC CpG islands are defined based upon hg19 coordinates), so the MIRA peaks were converted to hg19 coordinates (with a minimum re-mapping ratio of 0.95 without multiple output regions) using the lift-over function in Galaxy [59-61]. Likewise, the probe annotation file was converted to hg19 coordinates in the same way. Finally, peaks were annotated with UCSC CpG Islands showing at least 50% overlap (with either the peak or the island, including the 2 kb flanking region for the CpG island shores). The same DNA sample was used for the Illumina 450k array (GSE42707), which was processed in the same way as the HCT116 sample at the COH Functional Genomics Core.

Figure S1: Overview of COHCAP "Average by Site" Workflow

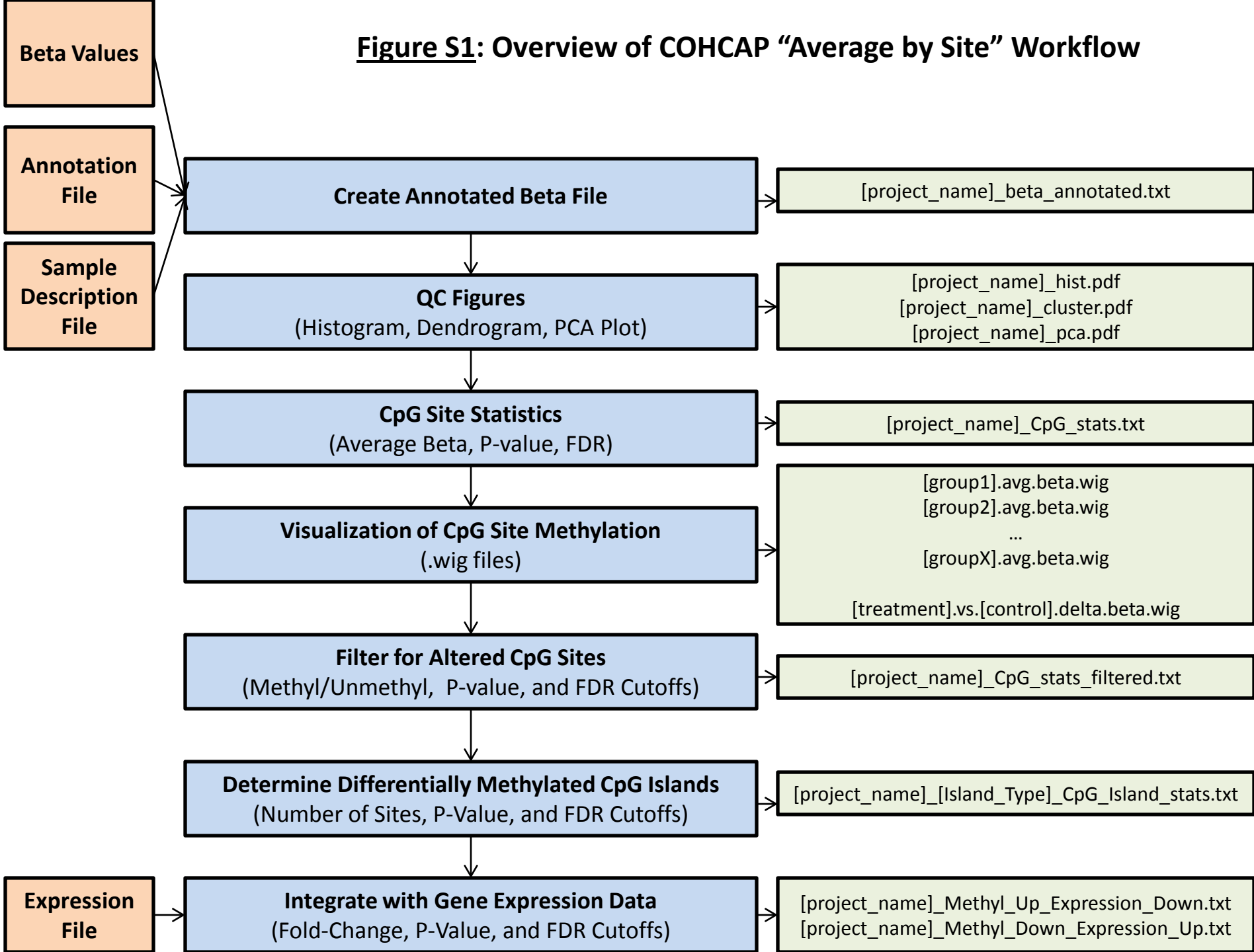


Figure S2: Overview of COHCAP "Average by Island" Workflow

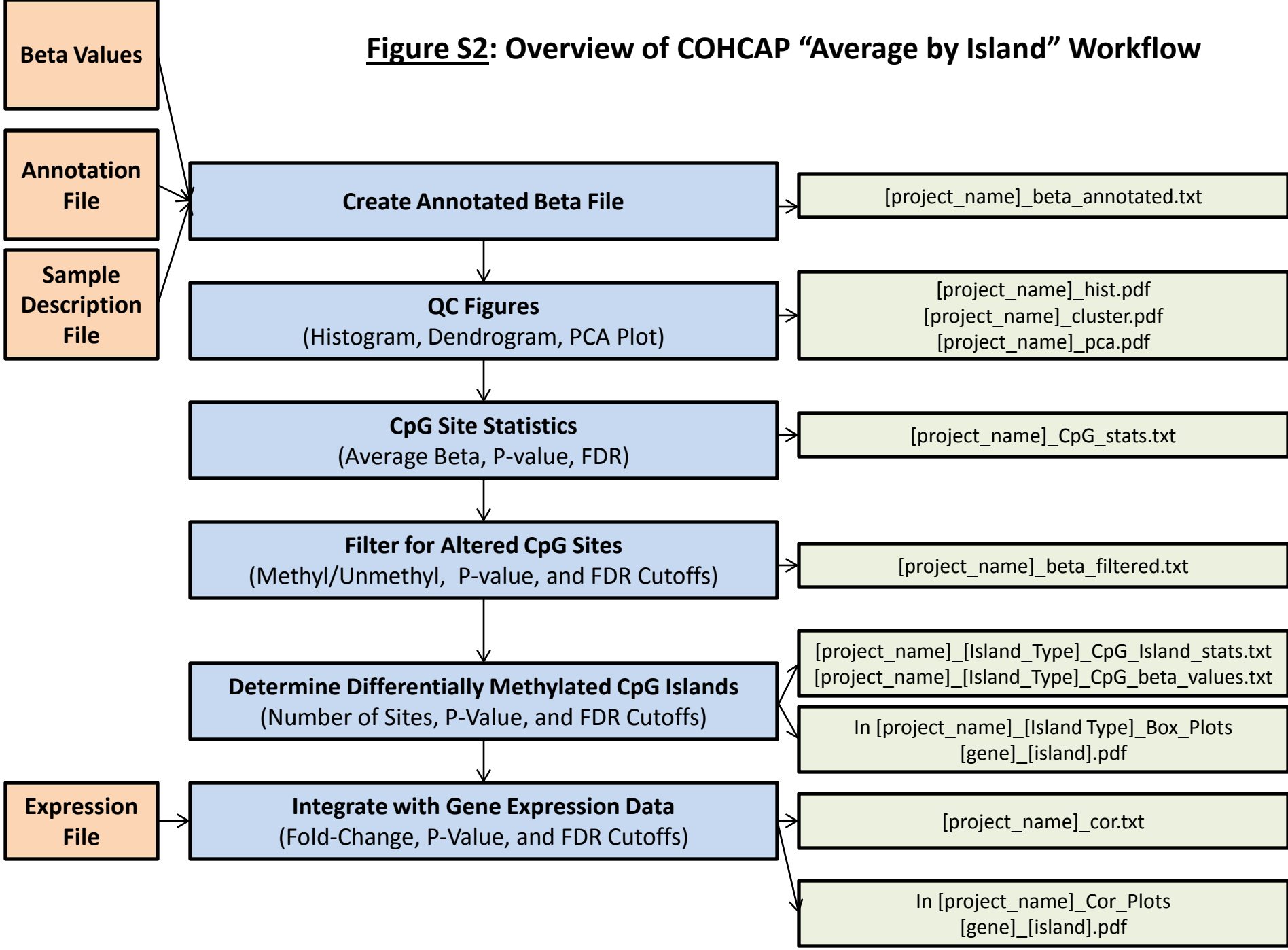


Figure S3: Overview of 450k Analysis in This Study

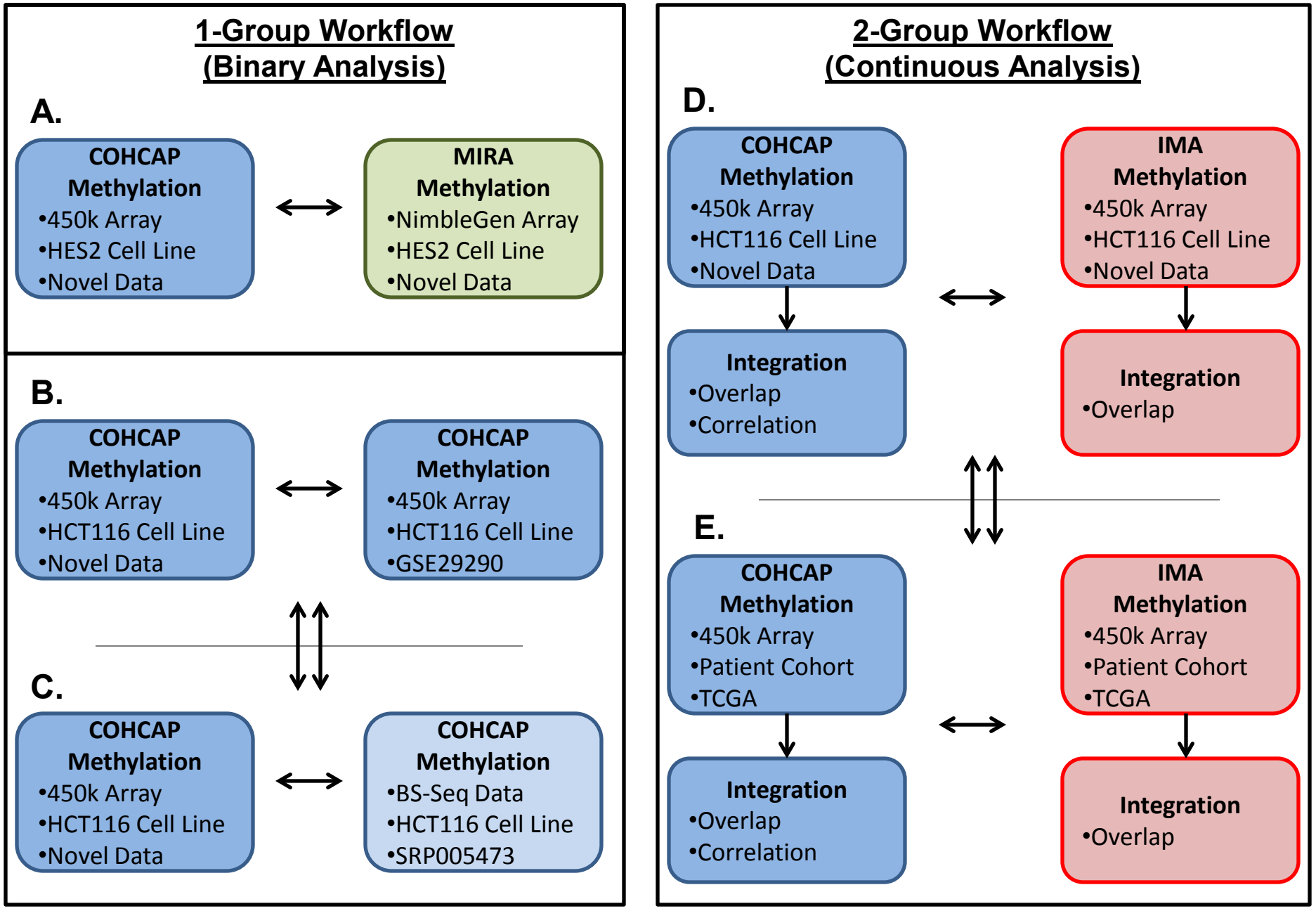


Figure S4: Overview of BS-Seq Analysis in This Study

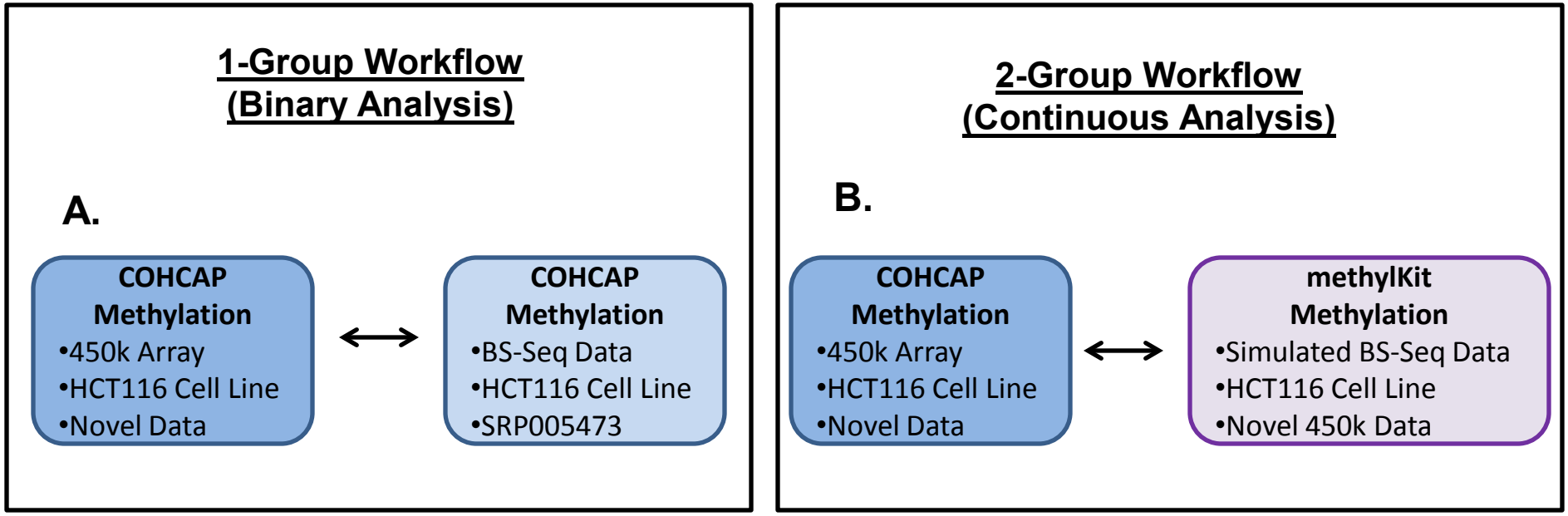
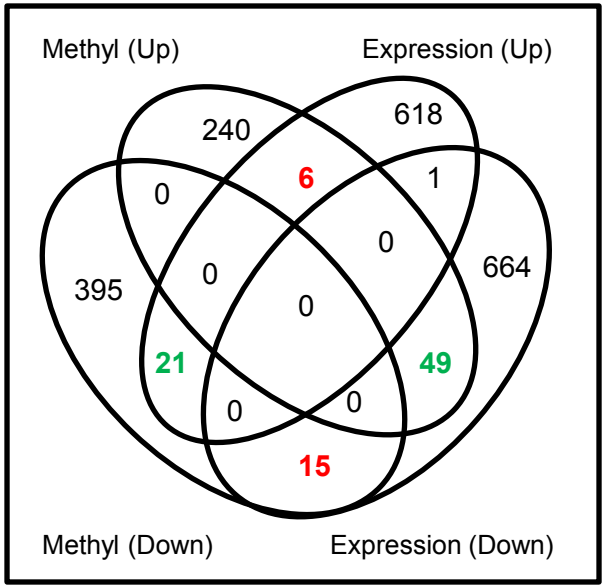
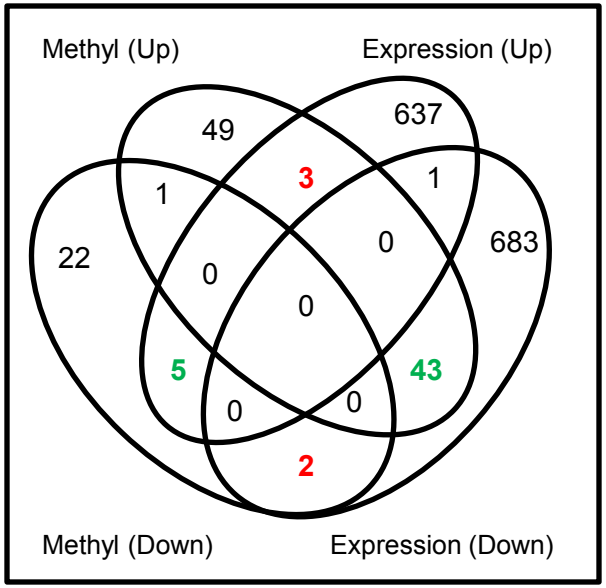


Figure S5: Overlap with Differential Gene Expression for the HCT116 Comparison

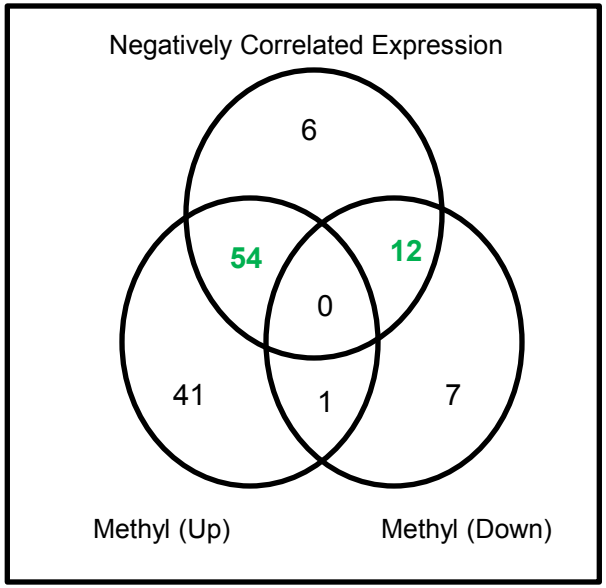
A. IMA



B. COHCAP (Overlap)



C. COHCAP (Correlation)



All overlap is defined based upon gene annotation. It is possible for a gene to show a negative correlation without meeting the criteria for differential methylation because the table of averaged beta values (used for correlation analysis) is only filtered based upon the number of significant CpG sites. Group-level averaging of beta values, p-values, and FDR values is calculated after this step.

Figure S6: Validation of Selected CpG Islands and Corresponding Gene Expression

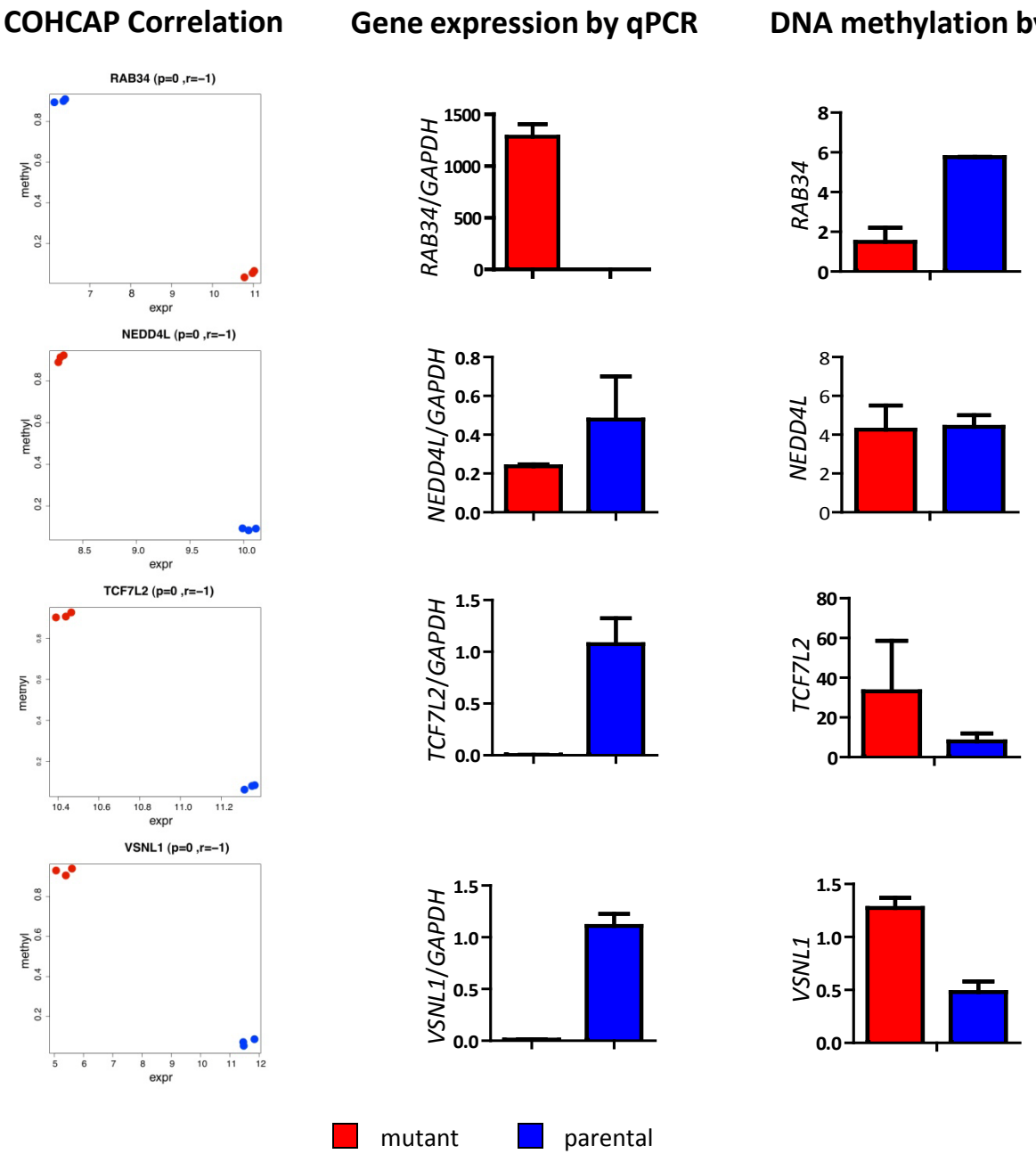
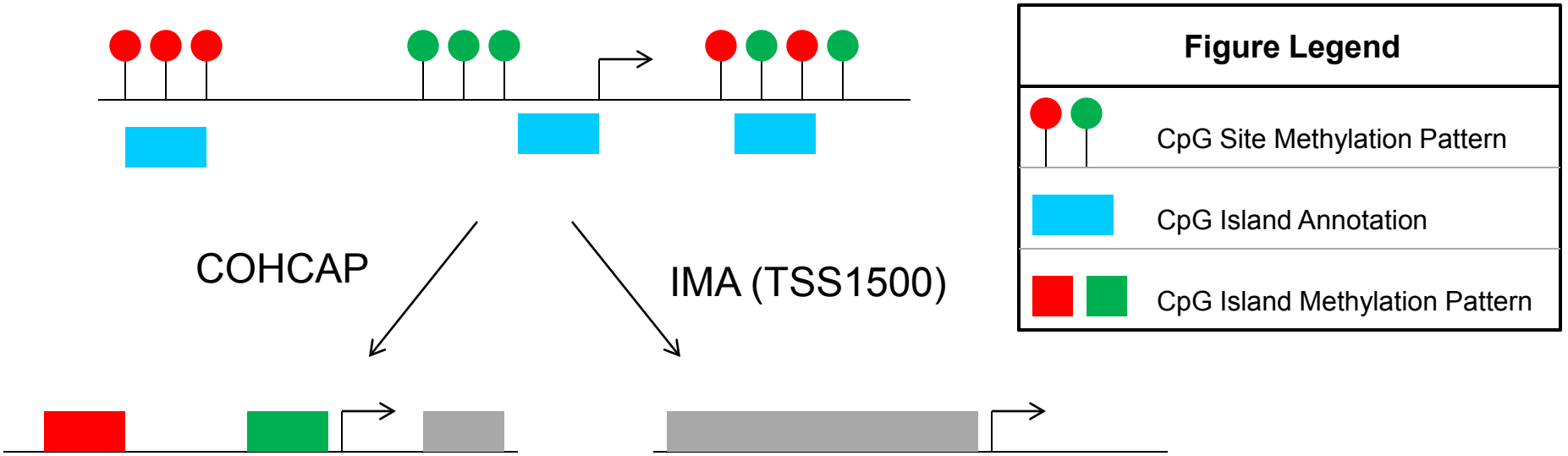


Figure S7: Visualization of CpG Sites in NEDD4L Promoter

Figure S8: Comparison of COHCAP and IMA Gene-Centric Analysis



Blue boxes represent CpG island annotations (in this paper, these are the UCSC CpG islands). Circles represent CpG sites (red for increased methylation, green for decreased methylation). Boxes between sites represent CpG island behavior (red for increased methylation, green for decreased methylation, and grey for unchanged methylation). Notice that a single promoter can contain multiple CpG islands and some CpG islands are not necessarily located upstream of the transcription start site (TSS). IMA also provides CpG island analysis but does not provide gene mappings. Also, IMA segregates the UCSC CpG islands into 5 regions (ISLAND, NSHORE, SSHORE, NSHELF, SSHELF), whereas COHCAP considers all of these regions to be a single functional unit . This is why the cluster of down-regulated CpG sites causes the 2nd CpG island to be considered up-regulated (e.g. the sites in the north shore are considered in the calculation of activity for the island).

Figure S9: Concordance with Gene Expression for the Mutant HCT116 Comparison

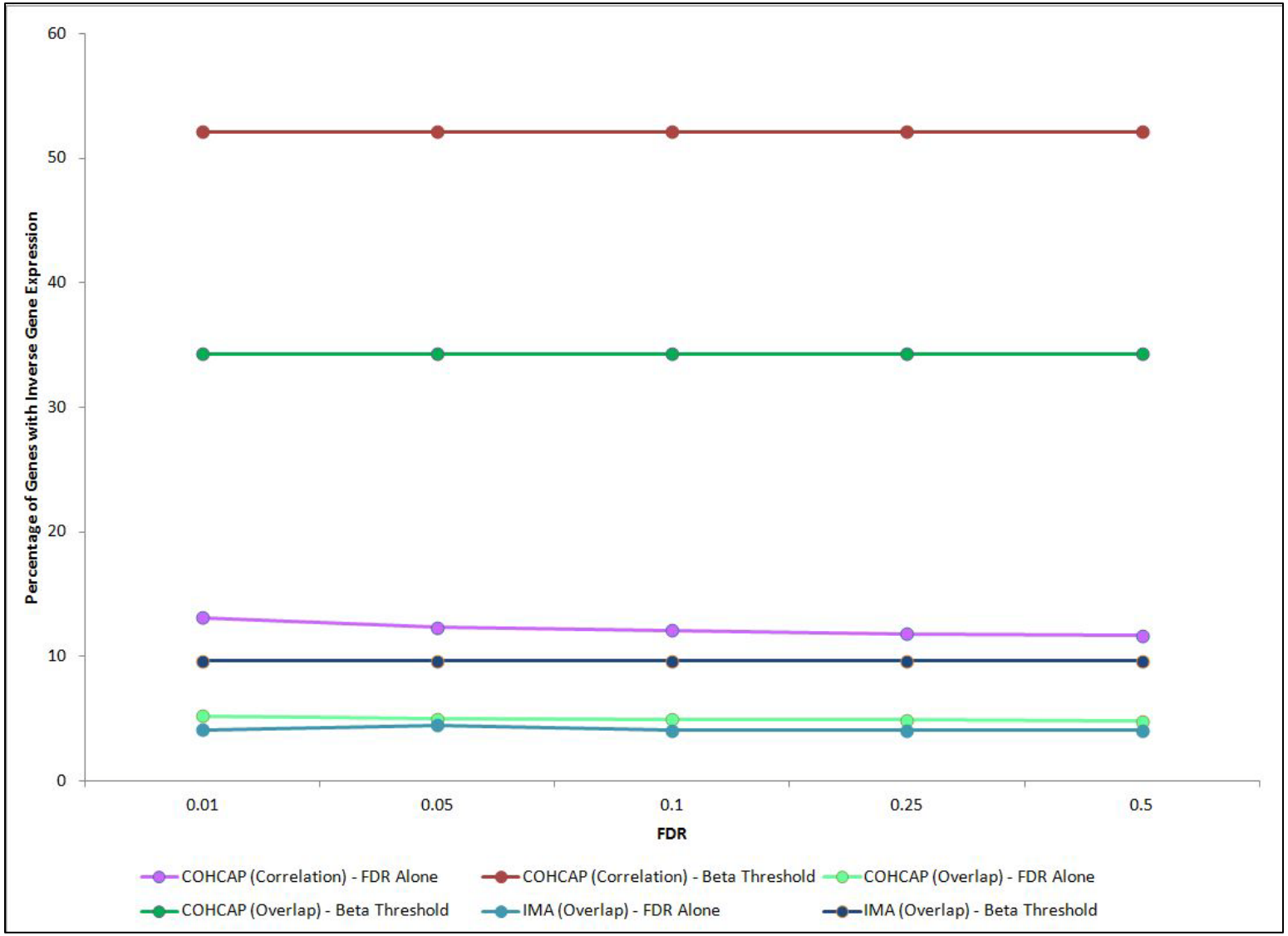


Figure S10: Integration via Correlation Outperforms Integration via Overlap

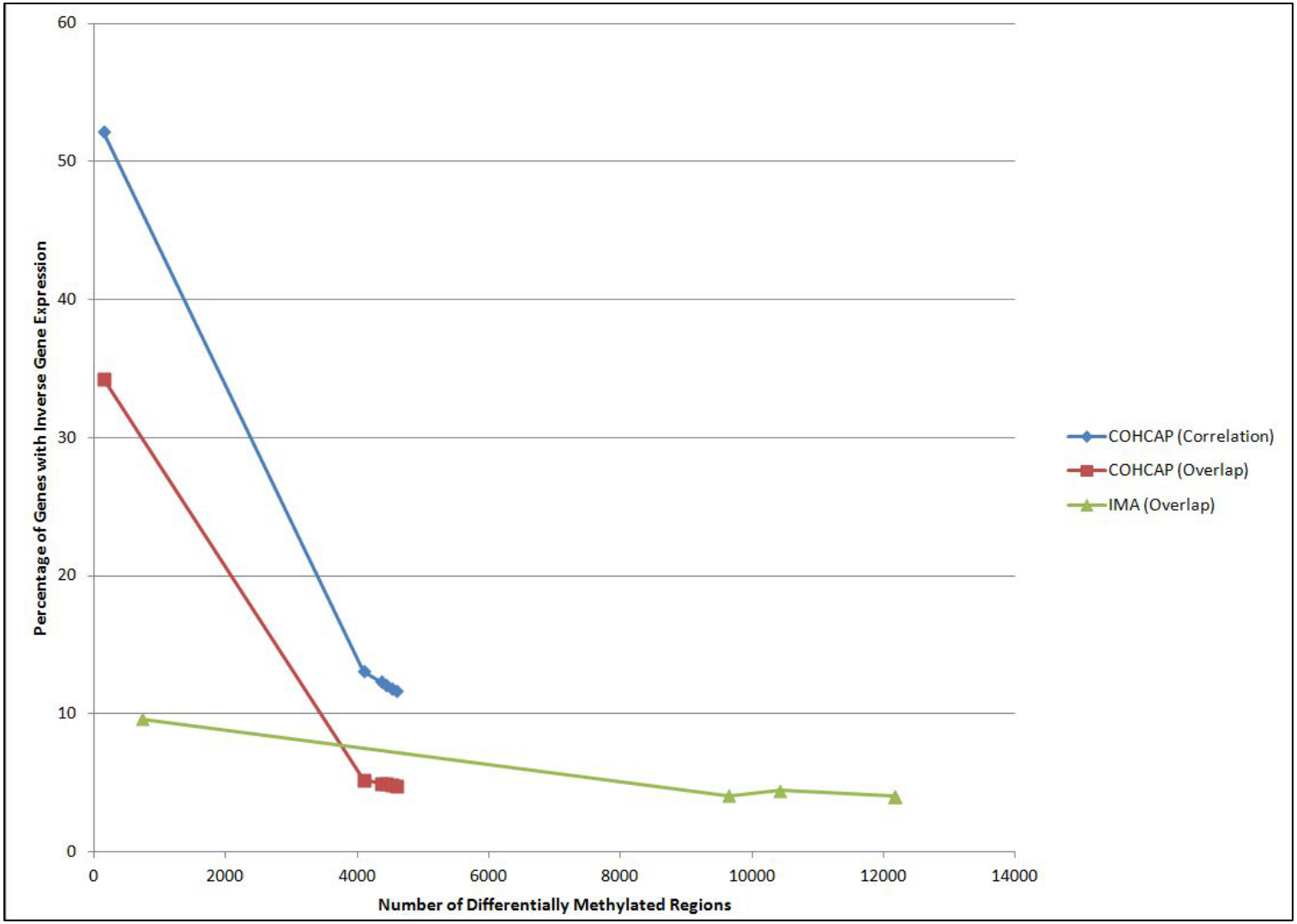
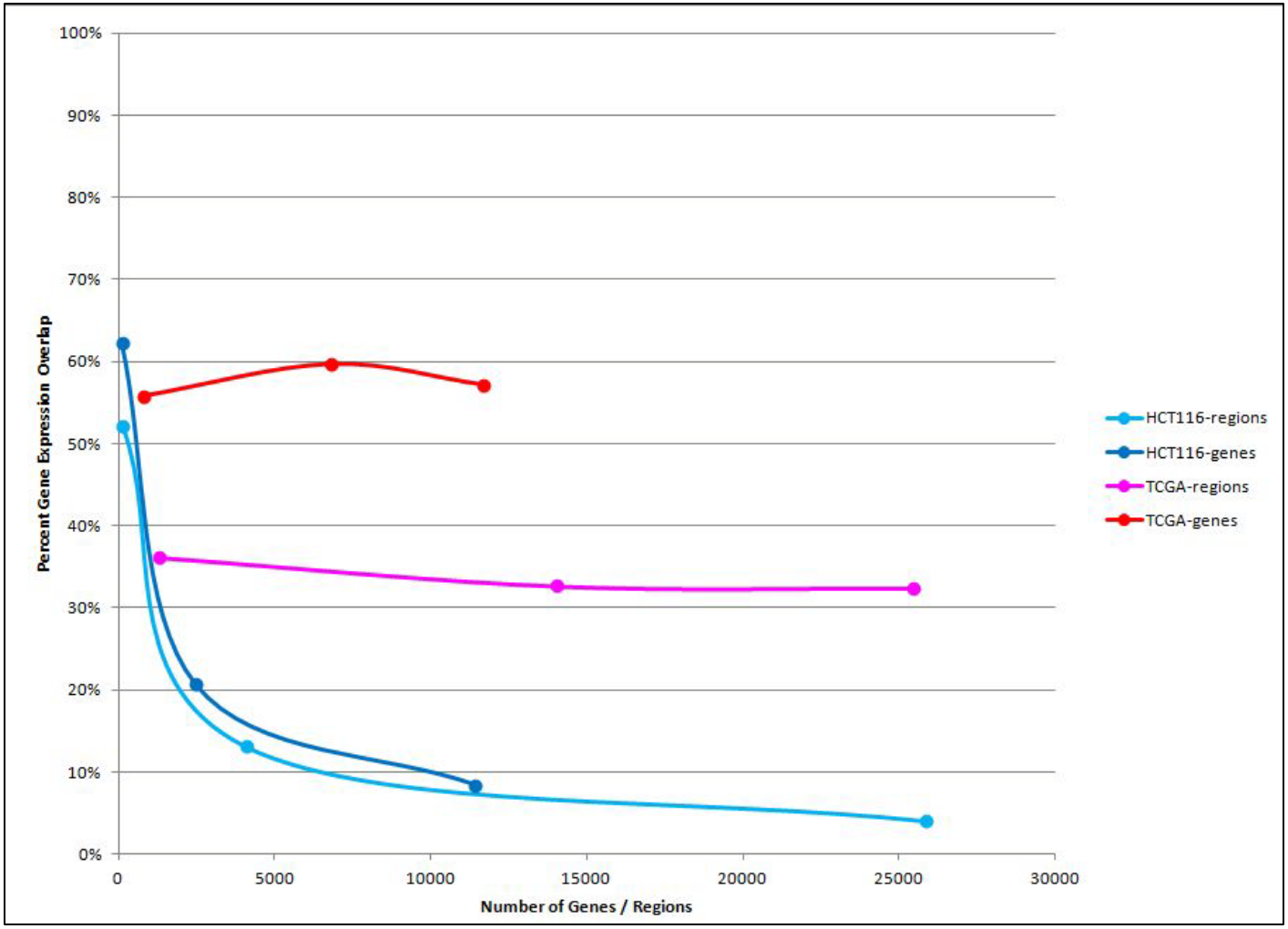


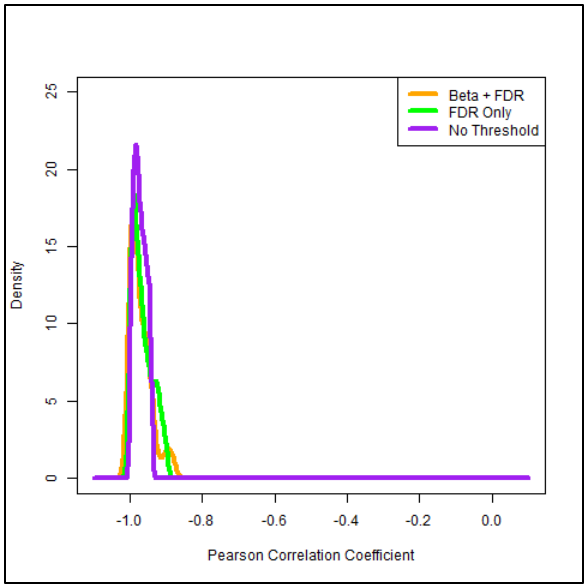
Figure S11: Gene Expression Concordance for Small and Large Gene Lists



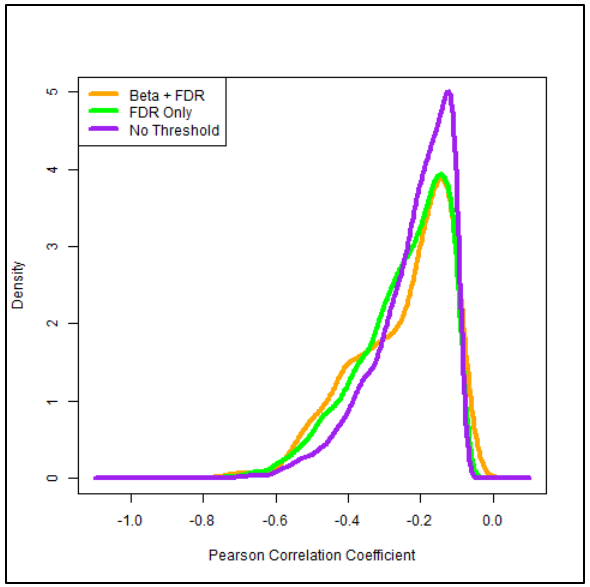
This scatter plot was created using COHCAP with 3 thresholds: 1) methylated / unmethylated threshold + FDR, 2) FDR alone, and 3) no threshold. Note that approximately 50% of the UCSC CpG islands do not map to genes.

Figure S12: Density Distribution for COHCAP Correlation Coefficients

HCT116



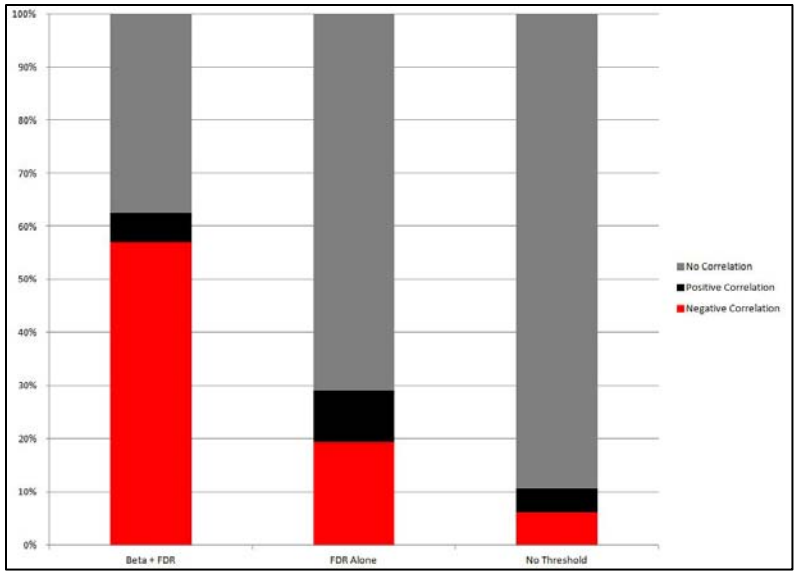
TCGA



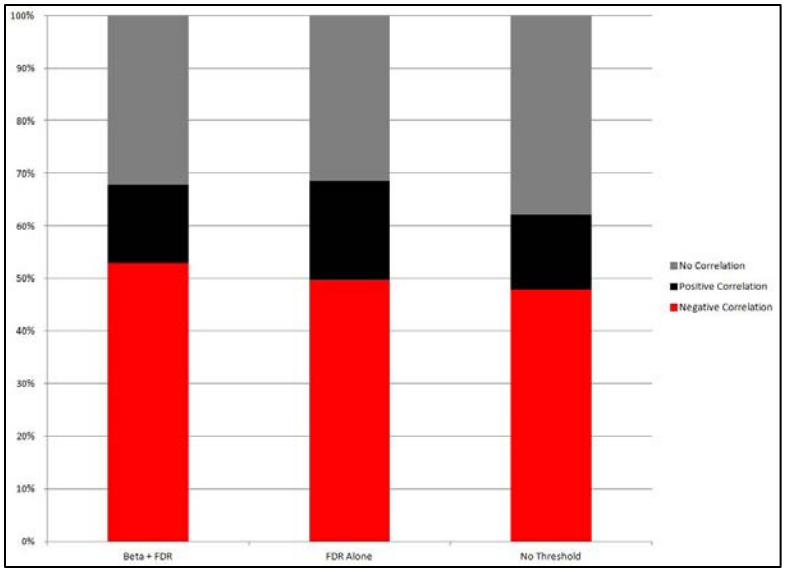
Density plots were created using COHCAP with 3 thresholds: 1) methylated / unmethylated threshold + FDR (orange lines), 2) FDR alone (green lines), and 3) no threshold (purple lines). Note that the correlation coefficient has to be quite strong to be detected in a small dataset but large patient cohorts can detect much more subtle correlations using the same criteria (and correlations tend to be weaker overall).

Figure S13: COHCAP Selects Regions Enriched for Negative Correlations

HCT116



TCGA



This scatter plot was created using COHCAP with 3 thresholds: 1) methylated / unmethylated threshold + FDR, 2) FDR alone, and 3) no threshold. Negative correlations are shown in red, positive correlations are shown in black, and non-significant correlations are shown in grey.

Figure S14: Overlap with Differential Gene Expression for the TCGA Comparison

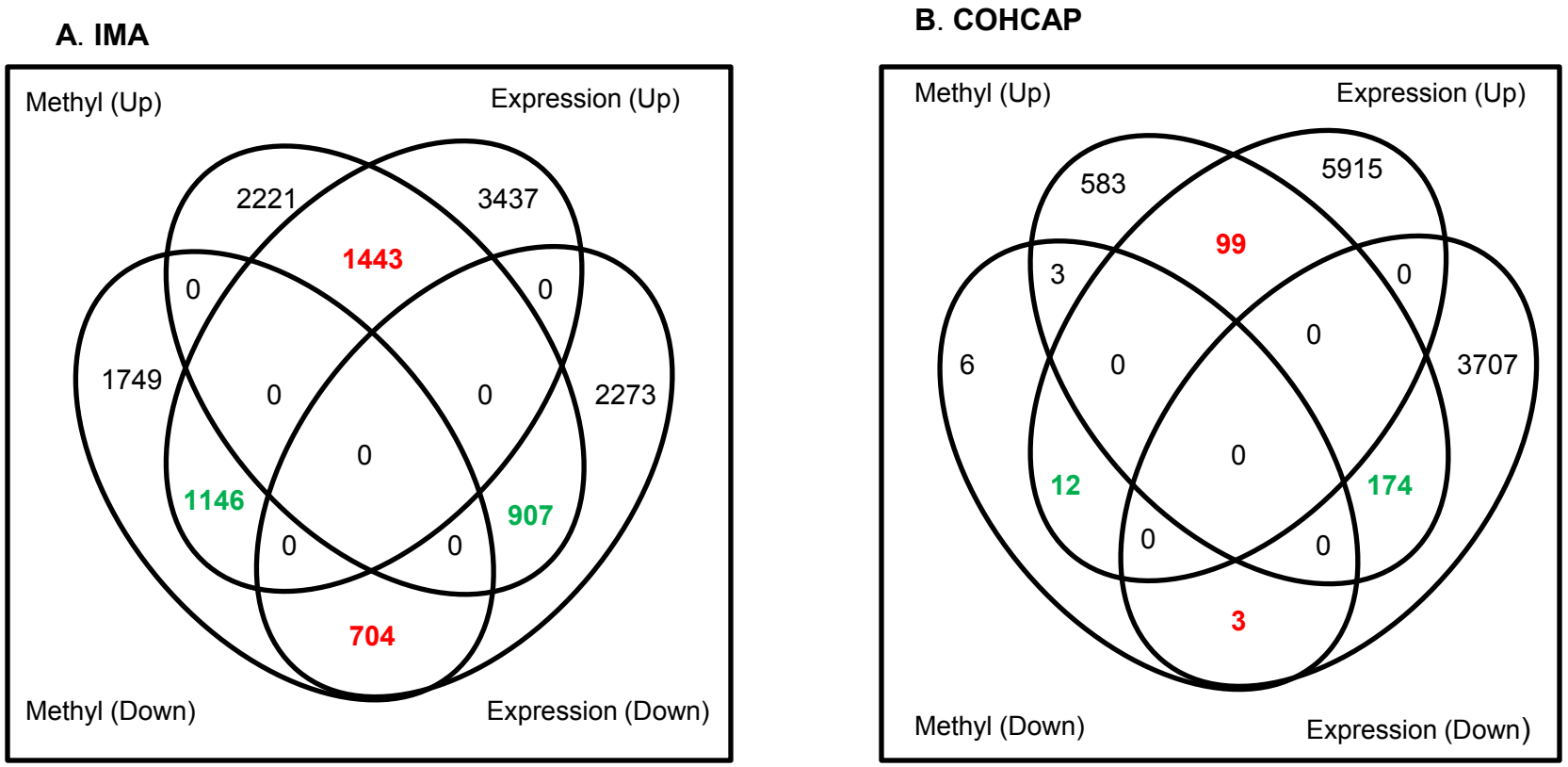
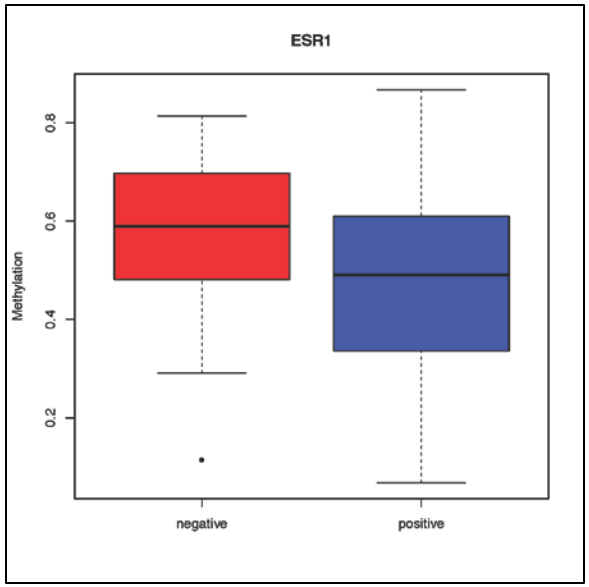
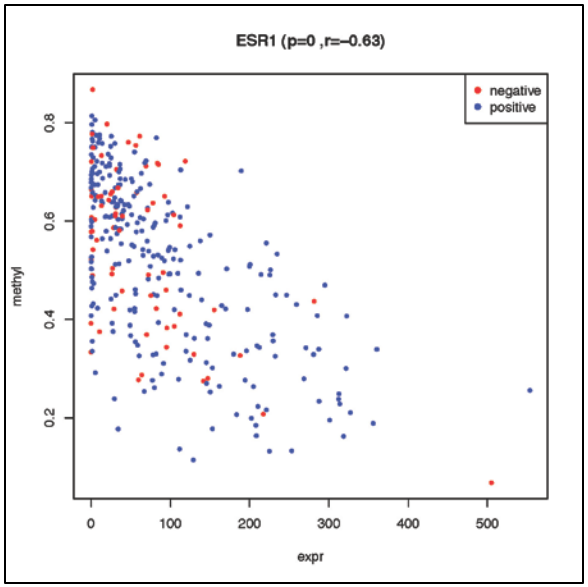


Figure S15: ER+ Patients Show Decreased Methylation in UCSC CpG Island





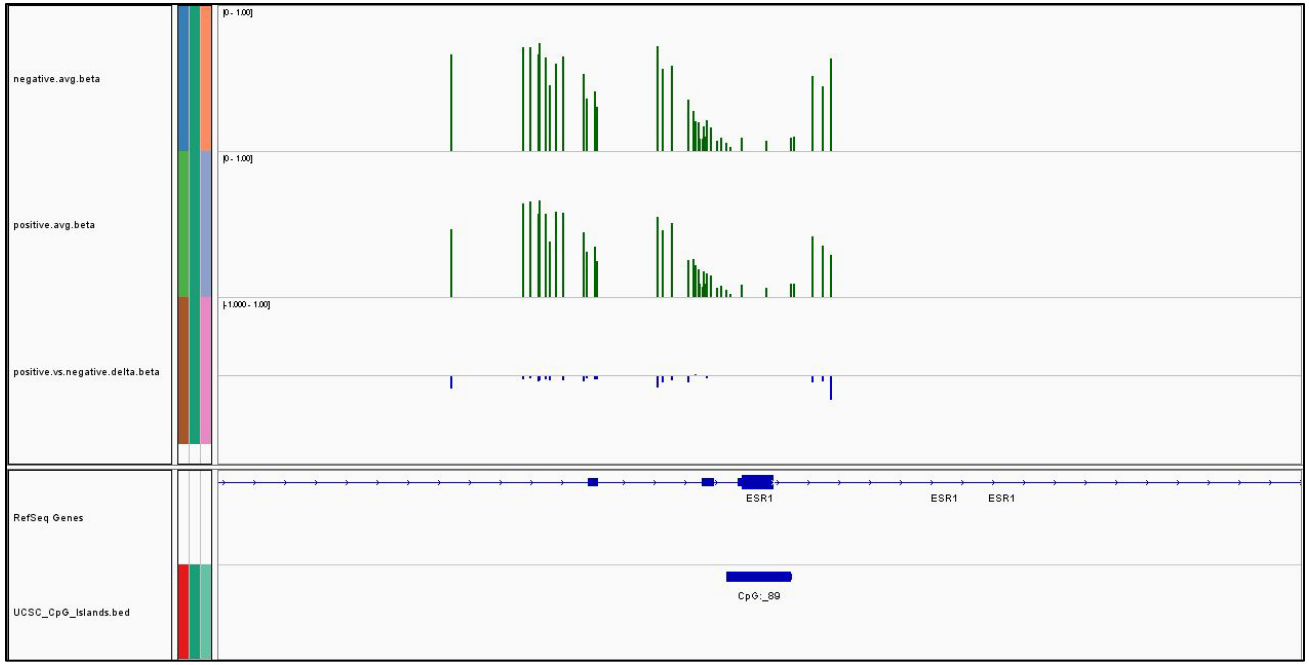


Figure S16: Overlap between TCGA Differentially Methylated Regions for IMA and COHCAP

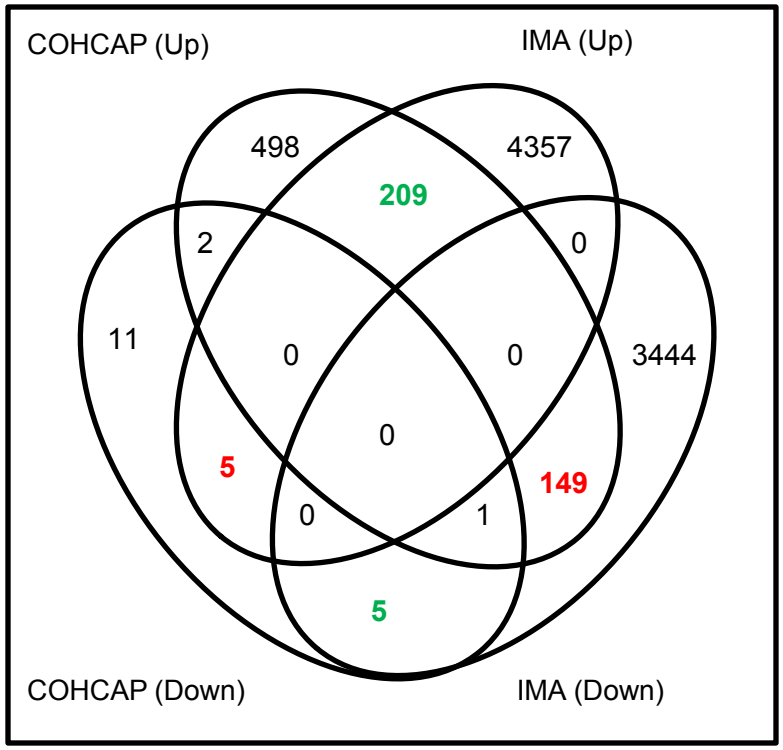


Figure S17: Differentially Methylated Regions Mapped to PLEC

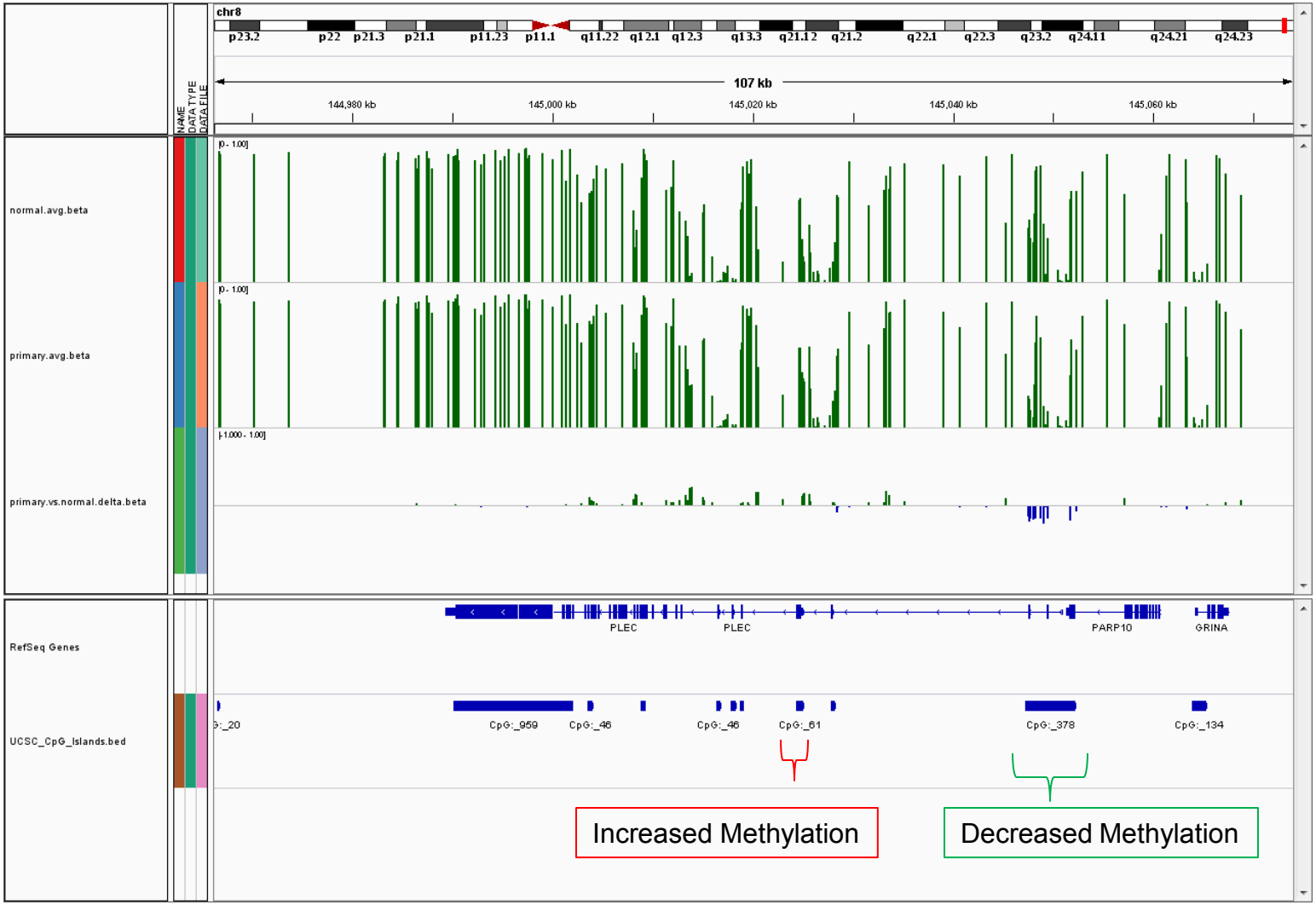


Figure S18: COHCAP Overlap with MIRA Peaks

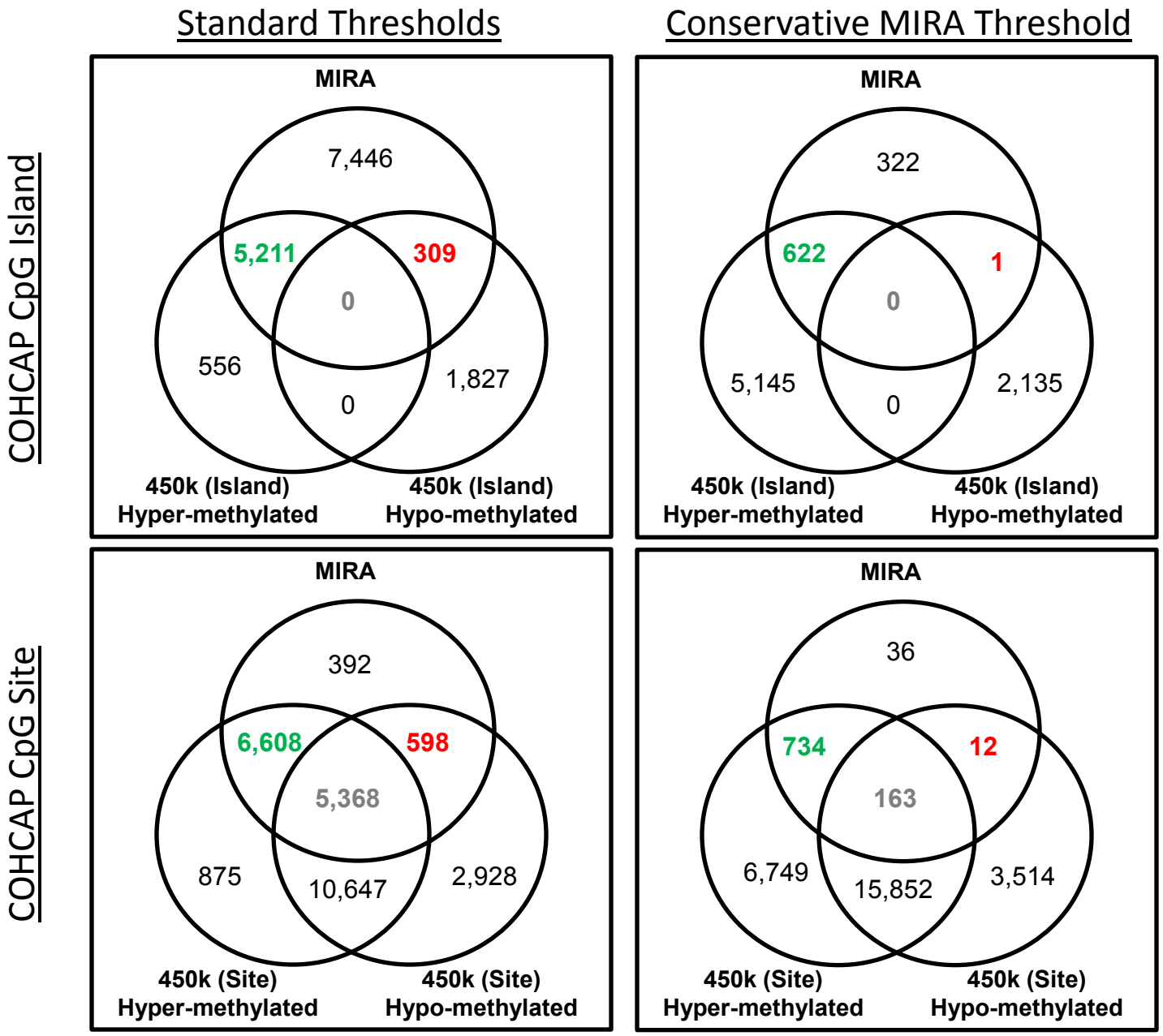


Figure S19: Visualization of Overlapping MIRA Peak

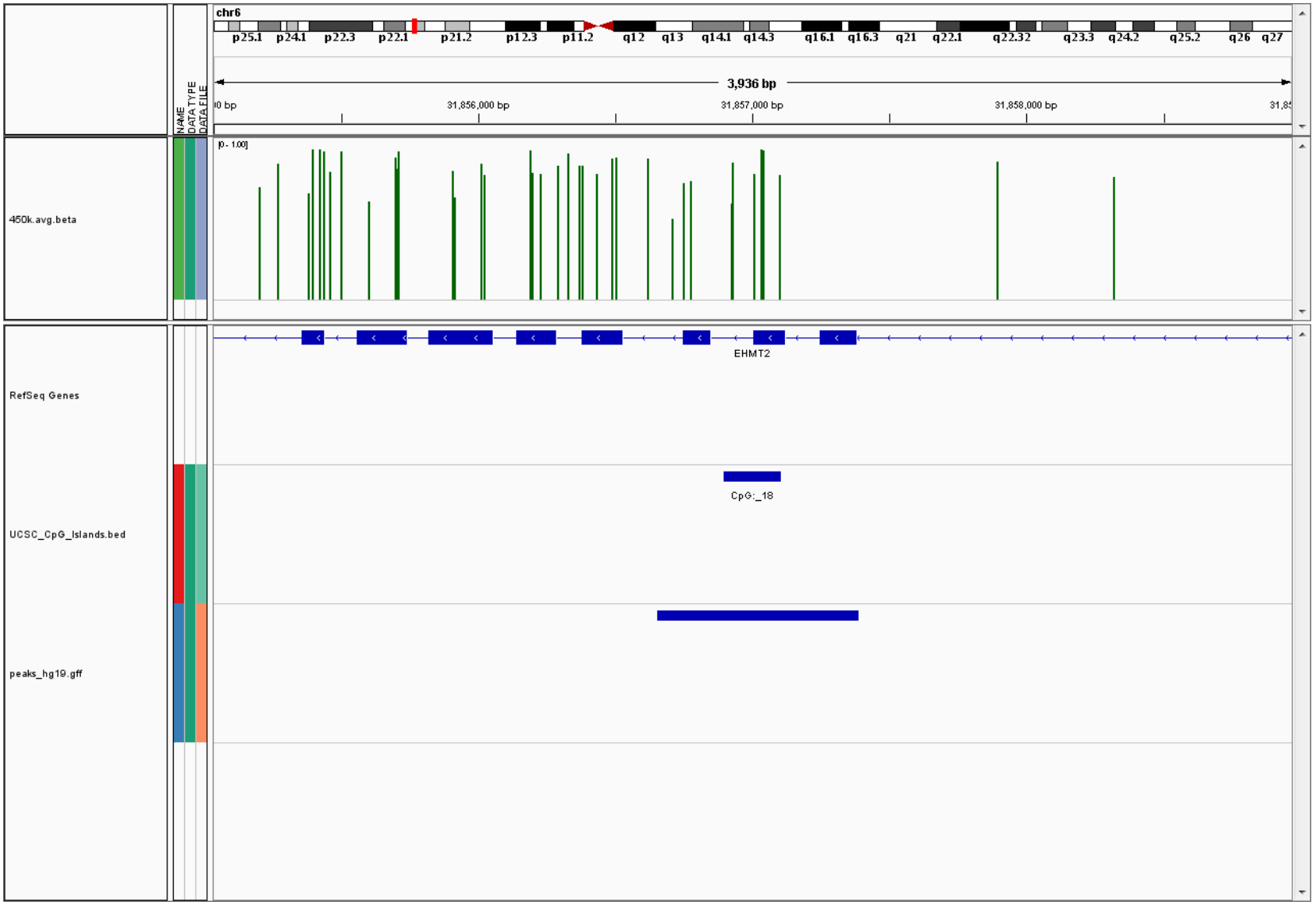
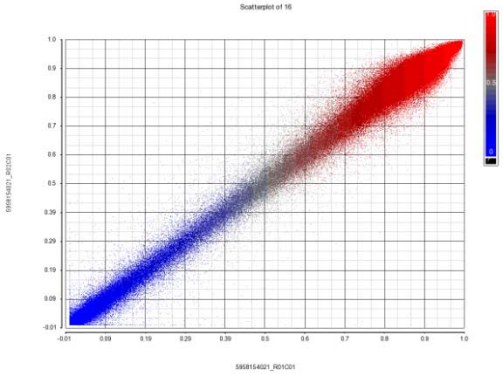


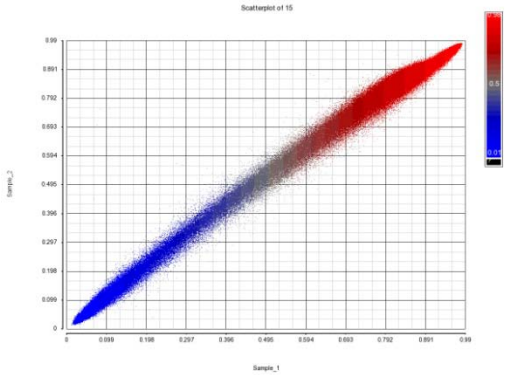
Figure S20: Correlation of CpG Site Signal for Biological Replicates

GSE29290
(Illumina Array, HCT116)



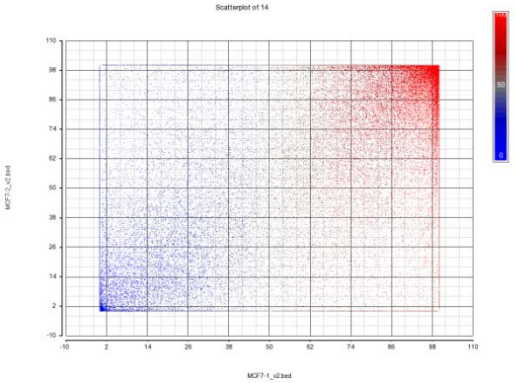
(ρ=1.00)

This Study
(Illumina Array, HCT116)



(ρ=1.00)

SRP005473
(BS-Seq, MCF7)



(ρ=1.00)

Figure S21: Decreased Methylation at the ESR1 TSS in Breast Tumors

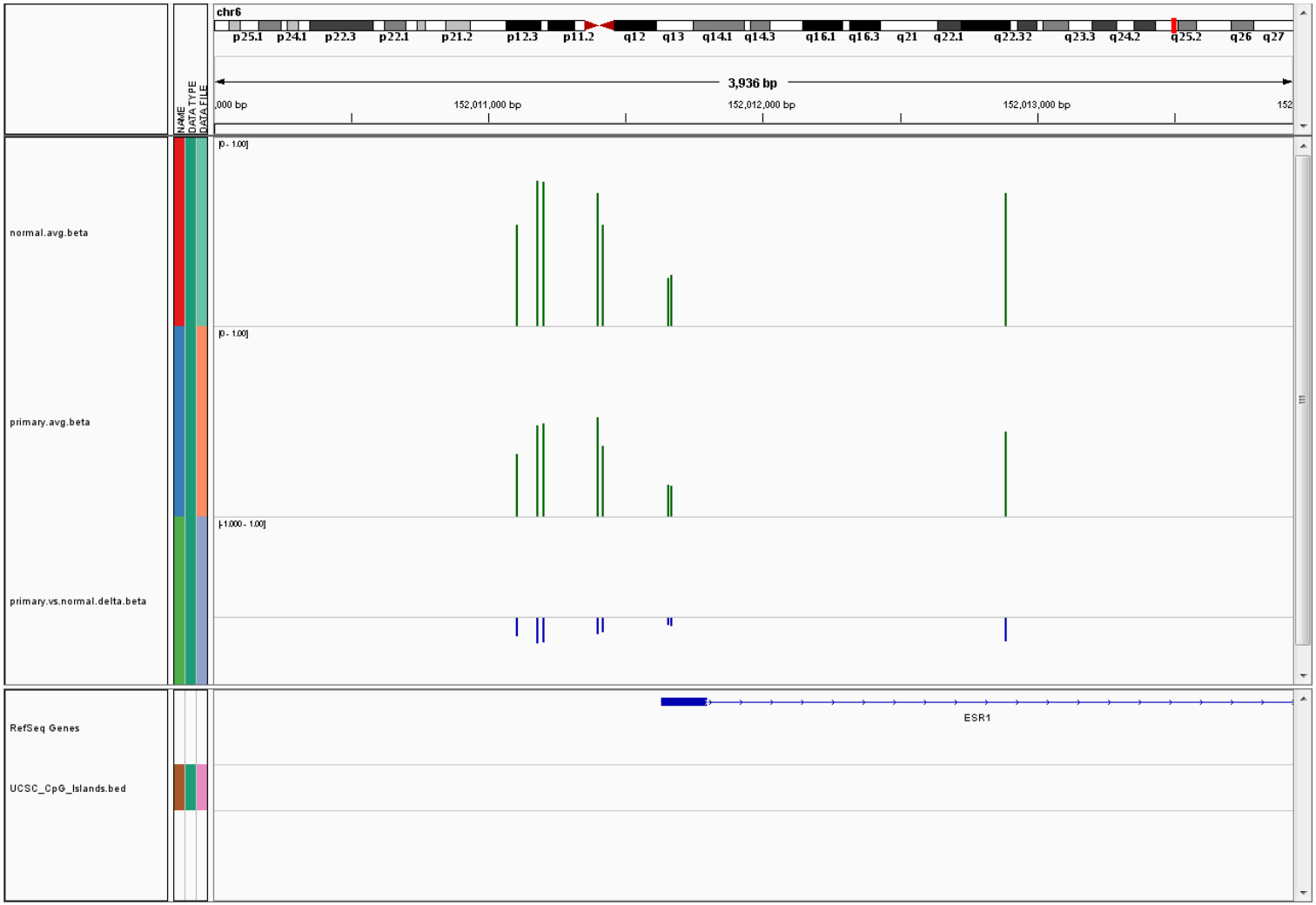


Table S1: Comparison of COHCAP to Other Algorithms

	COHCAP	IMA	CpGassoc	NIMBL		methylKit
Illumina Array	**Yes**	**Yes**	**Yes**	**Yes**		No
BS-Seq	**Yes**	No	No	No		**Yes**
CpG Site	**Yes**	**Yes**	**Yes**	**Yes**		**Yes**
CpG Island	**Yes (2 methods)**	**Yes (region)**	No	No		**Yes (windows)**
Criteria of Differential Methylation	• **Methyl Threshold** • **Unmethyl Threshold** • P-value • FDR	• **Delta-beta** • P-value • FDR	• P-value • FDR	• **Methyl Threshold** • **Unmethyl Threshold** • **Delta-beta** • P-value • FDR		• **Delta-beta** • P-value • FDR
1-Group Analysis	**Yes**	No	No	No		No
2-Group Analysis	**Yes**	**Yes**	**Yes**	**Yes**		**Yes**
>2 Group Analysis	**Yes**	**Yes (Continuous Variable)**	**Yes (Continuous Variable)**	No		No
Detection P-Value Cutoff	Optional	Required	Required	Required		NA
Visualization	• QC stats • **.wig file** • **Box-Plot** • **Scatterplot**	QC stats	• QC stats • **Scatterplot** • **Box-Plot**	• QC stats • **Gene View** • **Box-Plot** • **Venn Diagram**		• QC stats • **Bedgraph** • **Annotation plots**
Integration with Gene Expression	**Yes**	No	No	No		No
Programming Language	**Perl** **R** **(Java - GUI)**	**R**	**R** (MATLAB - GUI)	MATLAB		**R**

NOTE: COHCAP does provide delta beta values in the output file, but it is not currently considered as part of the filtering process for the workflows.

Table S2: Variations between COHCAP Algorithms

A. Average by CpG Site

	CpG Site Analysis	Filtering	CpG Island Analysis	Integration
1 Group	Unpaired: Average β	Unpaired: Average β	Unpaired: Fisher's Exact Test	NA
	Paired: NA	Paired: NA	Paired: NA	
2 Groups	Unpaired: Group Site β t-test	Methyl Cutoff Unmethyl Cutoff P-Value FDR	2-way ANOVA (Group, Site)	Overlap
	Paired: Group Site β 2-way ANOVA (Group, Pair)			
>2 Groups	Unpaired: Group Site β 1-way ANOVA (Group)	P-Value FDR	2-way ANOVA (Group, Site)	NA
	Paired: Group Site β 2-way ANOVA (Group, Pair)			

B. Average within CpG Island (Default)

	CpG Site Analysis	Filtering	CpG Island Analysis	Integration
1 Group	Unpaired: Average β	Unpaired: Average β	Unpaired: Average β	Correlation
	Paired: NA	Paired: NA	Paired: NA	
2 Groups	Unpaired: Group Site β t-test	Methyl Cutoff Unmethyl Cutoff P-Value FDR	Unpaired: Group Island β t-test	Correlation
	Paired: Group Site β 2-way ANOVA (Group, Pair)		Paired: Group Island β 2-way ANOVA (Group, Pair)	
>2 Groups	Unpaired: Group Site β 1-way ANOVA (Group)	P-Value FDR	Unpaired: Group Island β 1-way ANOVA (Group)	Correlation
	Paired: Group Site β 2-way ANOVA (Group, Pair)		Paired: Group Island β 2-way ANOVA (Group, Pair)	

*Method of Benjamini and Hochberg is always used for FDR calculation

Table S3: Relative Performance of CpG Island Algorithms on COH HCT116 Dataset

	Num of Regions	COHCAP Overlap (Avg by Island)	Expression Overlap	Expression Correlation
COHCAP (Avg by Island)	140 regions	125 genes (100%)	**48 genes (38.4%)**	**73 regions (52.1%)**
COHCAP (Avg by Site)	275 regions	115 genes (41.8%)	**32 genes (11.6%)**	NA
IMA (TSS200)	811 regions	70 genes (8.6%)	**15 genes (1.8%)**	NA
IMA (TSS1500)	726 regions	37 genes (5.1%)	**70 genes (9.6%)**	NA
IMA (ISLAND)	616 regions	**93 regions (15.1%)**	NA	NA
IMA (NSHELF)	933 regions	10 regions (1.1%)	NA	NA
IMA (NSHORE)	1,404 regions	48 regions (3.4%)	NA	NA
IMA (SSHELF)	880 regions	7 regions (0.8%)	NA	NA
IMA (SSHORE)	1,438 regions	44 regions (3.1%)	NA	NA

Table S4: Runtime for COH HCT116 Analysis (N=6)

Linux Server (Concurrent Usage, x64, CentOS Red Hat 4.1, 264 GB RAM, 6 x 2.27 GHz processors)

	COHCAP (Average by Site)	COHCAP (Average by Island)	IMA	methylKit
Annotate Beta File	0:00:22	0:00:22	NA	NA
QC Figures	0:00:42	0:00:43	NA	NA
Site Statistics	0:06:26	0:06:27	NA	NA
.wig file creation	0:00:12	NA	NA	NA
Filter Sites	0:00:04	0:00:08	NA	NA
Island Statistics	0:00:18	0:00:23	NA	NA
Integration	0:00:01	0.00:12	NA	NA
Total Time	0:08:05	0:08:15	0:05:58	0:38:38

Local PC (Minimal Concurrent Usage, x64, Windows 7, 24 GB RAM, 2 x 3.60 GHz processors)

	COHCAP (Average by Site)	COHCAP (Average by Island)	IMA	methylKit
Annotate Beta File	0:00:51	0:00:57	NA	NA
QC Figures	0:00:43	0:00:43	NA	NA
Site Statistics	0:05:58	0:05:20	NA	NA
.wig file creation	0:00:29	NA	NA	NA
Filter Sites	0:00:05	0:00:12	NA	NA
Island Statistics	0:00:14	0:00:02	NA	NA
Integration	0:00:01	0:00:02	NA	NA
Total Time	0:08:21	0:07:15	0:04:27	0:25:48

Local Mac (Minimal Concurrent Usage, x64, Mac OS 10.6, 4 GB RAM, 2 x 2.66 GHz processors)

	COHCAP (Average by Site)	COHCAP (Average by Island)	IMA	methylKit
Annotate Beta File	0:01:53	0:01:30	NA	NA
QC Figures	0:01:30	0:01:17	NA	NA
Site Statistics	0:05:59	0:05:19	NA	NA
.wig file creation	0:01:35	NA	NA	NA
Filter Sites	0:00:27	0:00:39	NA	NA
Island Statistics	0:00:30	0:00:02	NA	NA
Integration	0:00:01	0:00:02	NA	NA
Total Time	0:11:55	0:08:49	0:10:34	0:35:05

Table S5: "Average by Site" CpG Island Statistics for Validation Genes

Gene	Region	# Methyl	# Unmethyl	P-Value	FDR
RAB34	chr17:27044168-27045049	0	7	1.8e-11	1.6e-9
	chr17:27046856-27047273	0	4	4.8e-6	0.00012
NEDD4L	chr18:55862653-55862873	8	0	5.5e-12	5.8e-10
TCF7L2	chr10:114712115-114712544	4	0	4.6e-6	1.2e-4
VSNL1	chr2:17721537-17722021	7	0	3.7e-15	7.8e-13

"#Methyl" and "# Unmethyl" describe the number of CpG shows showing relative hyper- or hypo-methylation (respectively) between the mutant and parental HCT116 strains. Although two regions are listed for RAB34, they both corresponded to nearby UCSC CpG islands (with a single optimized coordinate of chr17:27,440,000-27,045,500). Likewise, validation for NEDD4L used an optimized coordinate of chr18:55,862,550-55,862,900 and TCF7L2 used the optimized coordinate of chr10:114,712,000-114,713,500. Visual inspection of the VSNL1 CpG island indicated that the original UCSC coordinates were sufficient. In general, visual inspection is necessary because the CpG shores are included in the CpG island mappings and the overall range will not provide the most precise location of the differentially methylated CpG sites. Genome coordinates are specified with respect to hg19.

Table S6: Relative Performance of COHCAP Workflows on TCGA Methylation Data

A. Comparison of COHCAP Workflows

	Avg by Site Patient IDs Paired	**Avg by Site Patient IDs All**	**Avg by Site All All**	**Avg by Island Patient IDs Paired**	**Avg by Island Patient IDs All**	**Avg by Island All All**
CpG Site	21,089 sites	20,935 sites	0 sites	11,973 sites	11,544 sites	0 sites
CpG Island	1,602 regions	1,298 regions	NA	1,289 regions	1,145 regions	NA
Integration	177 regions	150 regions	NA	252 regions	221 regions	NA

B. Comparison with IMA (Patient IDs, Paired Samples)

	Num of Regions	**COHCAP Overlap**	**Expression Overlap**	**Expression Correlation**
COHCAP (Avg by Island)	1,289 regions	880 genes (100%)	**186 genes (21.1%)**	**466 regions (36.2%)**
IMA (TSS200)	6232 genes	201 genes (3.2%)	**1512 genes (24.2%)**	NA
IMA (TSS1500)	8170 genes	214 genes (2.6%)	**2053 genes (25.1%)**	NA
IMA (ISLAND)	12051 regions	**1259 regions (10.4%)**	NA	NA

CpG site analysis is identical for either workflow, but site counts vary because the "Average by Island" workflow includes a minimum number of sites per island in this early filtering step (but this is not considered until the integration step for the "Average by Site" workflow). "Patient IDs" means 2-way ANOVA takes tumor vs. normal patient pairing consideration. "Paired" means only samples with paired tumor and normal data were included (N=134). "All" means all possible samples were included (N=562). Analysis was performed with methylated / unmethylated beta values = 0.3; p-value cutoff = FDR cutoff = 0.05; Minimum number of sites per island= 4; fold-change cutoff=1.5; correlation coefficient cutoff = -0.2. Although the sample size varies by ~4x, the number of regions identified is very similar with paired vs. all samples.

Table S7: Runtime for TCGA Analysis

A. Paired Samples, N=134

	COHCAP (Average by Site)	COHCAP (Average by Island)	IMA
Annotate Beta File	0:04:35	0:06:45	NA
QC Figures	5:18:30	5:55:55	NA
Site Statistics	6:41:20	6:13:03	NA
.wig file creation	0:00:12	NA	NA
Filter Sites	0:00:04	0:02:57	NA
Island Statistics	0:02:43	0:11:37	NA
Integration	0:00:57	0:02:18	NA
Total Time	12:08:21	12:32:35	0:18:50

B. All Samples, N=562

	COHCAP (Average by Site)	COHCAP (Average by Island)	IMA
Annotate Beta File	0:31:18	0:06:18	NA
QC Figures	11:18:09	11:31:52	NA
Site Statistics	25:35:41	24:04:06	NA
.wig file creation	0:00:13	NA	NA
Filter Sites	0:00:04	0:02:57	NA
Island Statistics	0:02:14	0:14:54	NA
Integration	0:01:00	0:03:05	NA
Total Time	37:28:39	36:03:12	1:06:43

Analysis was performed on a Linux Server (Concurrent Usage, x64, CentOS Red Hat 4.1, 264 GB RAM, 6 x 2.27 GHz processors). COHCAP and IMA both caused the less powerful local computers (either PC or Mac) to crash.

Table S8: Differentially Methylated Breast Cancer Biomarkers with Inverse Changes in Gene Expression

Gene	Methylation Status	Biomarker Application(s)	Identified with IMA?
ESR1	Decrease (-0.17)	Diagnosis Disease Progression Efficacy Prognosis	EXON1 (0.08)
GATA3	Decrease (-0.08)	Diagnosis	EXON1 (0.12) TSS200 (0.059)
C2orf40/ECRG4	**Increase (0.27)**	Disease Progression	**GENEBODY (0.071)** TSS200 (-0.032)
CCND2	**Increase (0.26, 0.31)**	Efficacy	GENEBODY (-0.044) **TSS200 (0.026)**
GLI2	**Increase (0.34)**	Efficacy	**EXON1 (0.043)** **GENEBODY (0.020)** **TSS200 (0.043)** TSS1500 (-0.040)
GSTM2	**Increase (0.20)**	Diagnosis Prognosis	GENEBODY (-0.11) **TSS1500 (0.11)**
IGFBP3	**Increase (0.15)**	Diagnosis Efficacy	**EXON1 (0.072)** **GENEBODY (0.030)** **TSS1500 (0.12)**
LAMP3	**Increase (0.17)**	Diagnosis	**TSS1500 (0.076)**
PDPN	Increase (0.29)	Diagnosis	[None]
PROM1	**Increase (0.25)**	Efficacy	**EXON1 (0.10)** **TSS1500 (0.048)**
PTGS2	Increase (0.25)	Diagnosis Efficacy Prognosis	[None]
SFRP1	**Increase (0.19)**	Prognosis	**TSS1500 (0.026)**

*Only TSS200, TSS1500, EXON1, and GENEBODY IMA regions have mappings to genes are considered in the "Identified with IMA?" column. Delta beta values are shown in parentheses (to check concordance of methylation change). Genes with similar trends are shown in bold. IGFBP3 and GSTM2 are shown in red because they are the only genes with roughly corresponding magnitude of change and location of CpG Island.

Table S9: Selected IMA Breast Cancer Biomarkers with Inverse Overlap in Gene Expression

Gene	Methylation Status	Biomarker Application(s)	Identified with COHCAP?
BIRC5	Decrease (-0.07)	Disease Progression Prognosis	No
BRCA1	Decrease (-0.01)	Disease Progression Prognosis	No
BRCA2	Decrease (-0.09)	Disease Progression Prognosis	No
CD44	Decrease (-0.03)	Disease Progression Prognosis	No
ENAH	Decrease (-0.02)	Disease Progression	No
FOXA1	Decrease (-0.07)	Disease Progression Prognosis	No
PCGF2	Decrease (-0.07)	Prognosis	No
RPS6KB2	Decrease (-0.17)	Prognosis	No
SLC2A1	Decrease (-0.08)	Prognosis	No
ALDH1A1	Increase (0.47)	Disease Progression Prognosis	No
APC	Increase (0.08)	Disease Progression Prognosis	No
ARHGAP19	Increase (0.06)	Disease Progression	No
CEBPA	Increase (0.06)	Disease Progression Prognosis	No
CRMP1	**Increase (0.04)**	Disease Progression	**Yes (0.28)**
DPYD	Increase (0.10)	Disease Progression Prognosis	No
ENG	Increase (0.07)	Disease Progression Prognosis	No
ESR2	Increase (0.06)	Disease Progression Prognosis	No
FERMT2	Increase (0.09)	Prognosis	No
GSTM2	**Increase (0.11)**	Prognosis	**Yes (0.20)**
GSTM5	Increase (0.05)	Prognosis	No
PXN	Increase (0.03)	Prognosis	No

IMA statistics for TSS1500 regions are considered for this analysis, and delta beta values are shown in parentheses. IMA identified 33 biomarkers with decreased methylation and increased expression and 22 biomarkers with increased methylation and decreased expression. The subset of genes associated with disease progression and prognosis (instead of disease progression, diagnosis, efficacy, prognosis, or response to therapy) are shown in the table above. A total of 2039 genes (2.7% known biomarkers) were considered for IMA biomarker analysis while 247 genes were considered for COHCAP biomarker analysis (4.9% known biomarkers).

Table S10: Previously Reported Breast Cancer Biomarkers Influenced by Gene Expression

Gene	Delta Beta	Methylation P-value	Correlation Coefficient	Correlation P-value	Previous Publication	IPA Biomarker?
ESR1	-0.17	1.4×10^{-18}	-0.36	2.4×10^{-17}	Sproul et al. 2011 [44]	**Yes**
C2orf40/ECRG4	0.27	2.4×10^{-19}	-0.32	5.5×10^{-15}	Sabatier et al. 2011 [62]	**Yes**
CCND2	0.29	5.8×10^{-21}	-0.21	4.3×10^{-6}	How Kit et al. 2012 [63]	**Yes**
PCDH10	0.22	1.7×10^{-16}	-0.08	0.047	Faryna et al. 2012 [64]	No
POU4F1	0.29	2.2×10^{-22}	-0.13	0.0025	Faryna et al. 2012 [64]	No
SIM1	0.33	2.1×10^{-23}	-0.10	0.024	Faryna et al. 2012 [64]	No
TAC1	-0.32	1.7×10^{-21}	-0.11	0.0098	Faryna et al. 2012 [64]	No

CCND1 averaged between multiple CpG islands that map to the same gene.

Table S11: Top 10 COHCAP 450k Methylated CpG Islands for MIRA Comparison

A. CpG Island Analysis

COHCAP CpG Island	Gene	MIRA Peak?	3 x 720k Coverage?
chr6:31856896-31857104	EHMT2	Yes	Yes
chr6:31856896-31857104	BAT2	Yes	Yes
chr6:32063533-32065044	TNXB	Yes	Yes
chr6:32134460-32135196	PPT2	Yes	Yes
chr6:32046815-32047094	TNXB	Yes	Yes
chr6:31691425-31691718	LY6G6C	Yes	Yes
chr14:101531643-101532384	MIR377	Yes	Yes
chr6:32729342-32729877	HLA-DQB2	Yes	Yes
chr6:33288733-33289008	ZBTB22;DAXX	Yes	Yes
chr16:89345463-89348521	ANKRD11	Yes	Yes

B. CpG Site Analysis

Probe	COHCAP CpG Island	Gene	MIRA Peak?	3 x 720k Coverage?
cg20741134	chr1:181382290-181382848	NA	No	No
cg14377739	chr9:139701953-139703088	KIAA1984;LOC100131193	No	Yes
cg15027907	chr19:48076461-48076877	NA	No	Yes
cg12028762	chr19:3500416-3500840	DOHH	No	Yes
cg14462432	chr9:127572126-127572641	OLFML2A	Yes	Yes
cg19968916	chr10:135089969-135090491	ADAM8	No	Yes
cg06535952	chr19:56125349-56128167	NA	No	No
cg00710721	chr16:1179618-1179942	NA	No	Yes
cg17279652	chr5:180622178-180622658	TRIM7	Yes	Yes
cg24161106	chr17:17715696-17716219	MIR33B;SREBF1	No	Yes

Table S12: Runtime for Single-Group Analysis

	Linux Server	Local PC	Local Mac
OS	CentOS Red Hat 4.1	Windows 7	Mac OS 10.6
RAM	264 GB	24 GB	4 GB RAM
Processors	6 x 2.27 GHz	2 x 3.60 GHz	2 x 2.66 GHz
Annotate Beta File	0:00:24	0:00:49	0:01:44
QC Figures	0:00:43	0:00:44	0:01:17
Site Statistics	0:01:06	0:02:23	0:01:40
.wig file creation	0:00:04	0:00:09	0:00:20
Filter Sites	0:00:05	0:00:20	0:00:30
Island Statistics	4:41:00	3:13:48	5:01:42
Total Time	4:43:22	3:18:13	5:07:13

The HCT116 data produced for this study was used for all of these benchmarks. The "Average by Site" algorithm was used in COHCAP. The Linux server had normal concurrent usage but the local machines had minimal concurrent usage.

Table S13: Simulated Overlap of COHCAP CpG Islands in BS-Seq Data using methylKit

Simulation	COHCAP Recovery	# of methylKit Regions
100x Coverage	139 / 140 (99.3%)	221 / 14,449 windows
10x Coverage	138 / 140 (98.6%)	220 / 12,174 windows
5x Coverage	138 / 140 (98.6%)	213 / 8,940 windows
Vary Coverage per Site		
10x±5x Coverage	140 / 140 (100%)	218 / 11,843 windows
Conservative Windows (methylKit FDR < 0.001)		
10x Coverage	137 / 140 (97.6%)	215 / 7,290 windows
5x Coverage	133 / 140 (95.0%)	193 / 2,803 windows
Conservative Windows (methylKit FDR < 1e-6)		
10x Coverage	134 / 140 (95.7%)	199 / 3,236 windows
5x Coverage	126 / 140 (90%)	159 / 800 windows
Conservative Windows (methylKit FDR < 1e-10)		
10x Coverage	129 / 140 (92.1%)	172 / 1,118 windows
5x Coverage	114 / 140 (81.4%)	130 / 369 windows

In most cases, methylKit regions are defined using as showing at least a 30% methylation difference and an FDR < 0.05, except for the “conservative windows” which are filtered for a much lower FDR (to see if COHCAP regions ranked highly in the list of methylKit regions).