Supplemental Material

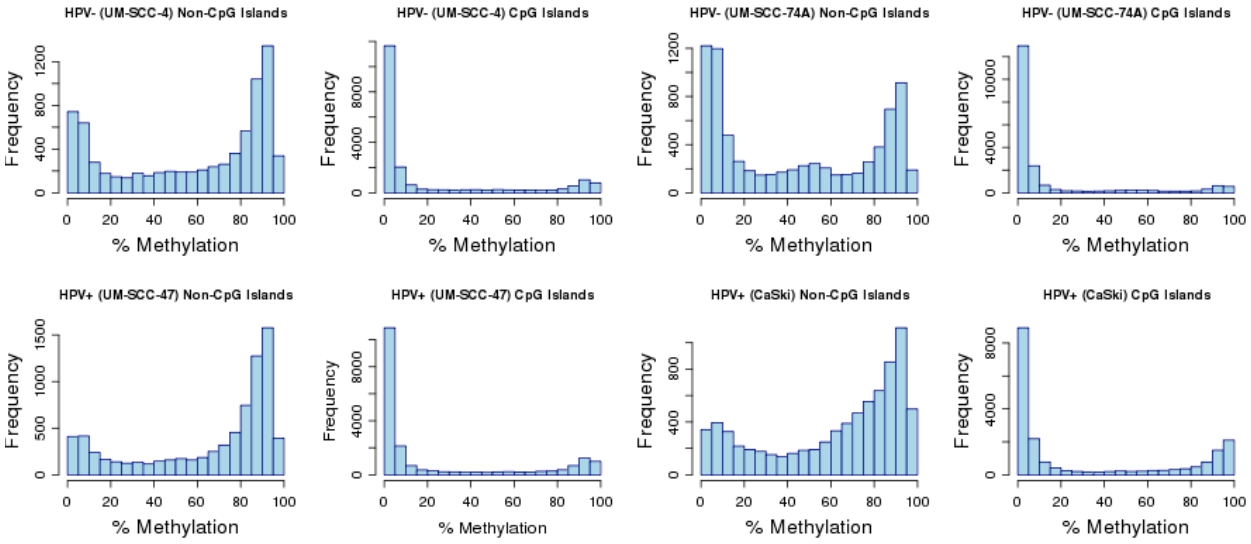


Figure S1. Relative CpG island and non-CpG island methylation in HPV(+) and HPV(-) tumor lines. The higher DNA methylation in HPV(+) tumor cells compared to HPV(-) tumor cells was observed both in Non-CpG island regions and CpG islands.

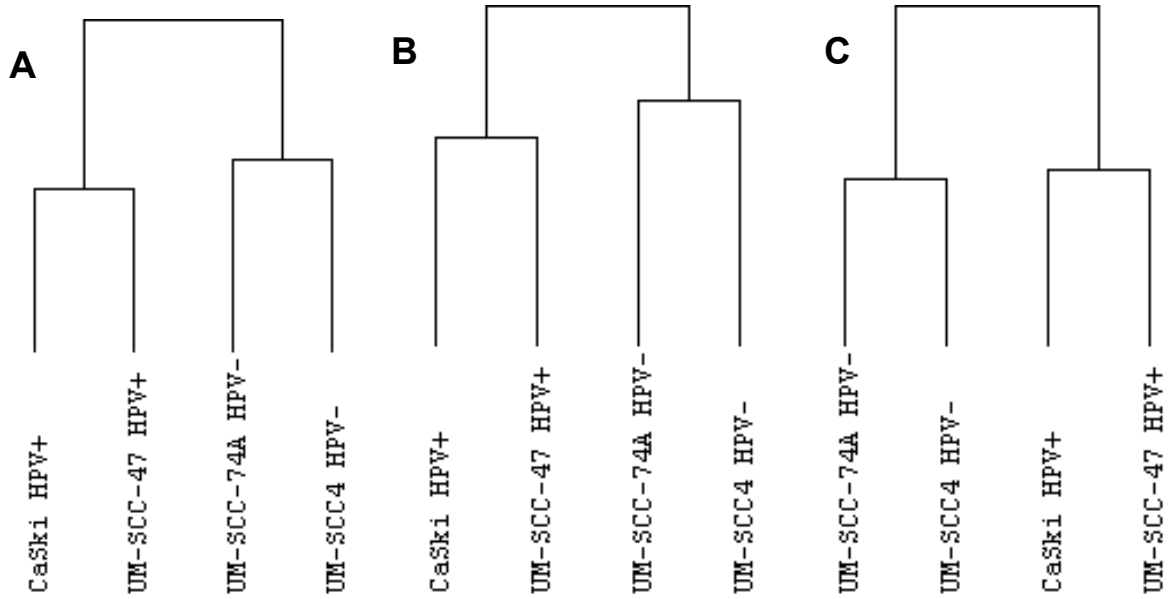


Figure S2. Unsupervised hierarchical clustering of the Illumina Infinium data, filtered based on three different levels of standard deviation of beta values across all four cell lines (see methods), was able to clearly separate the samples according to their HPV status. CpG sites with low standard deviation for beta values across all four cell lines were filtered out, which is a common approach to remove genes that are not informative, and values for each site were centered by their mean. Hierarchical clustering with correlation similarity measures and average linkage clustering was performed using Gene Cluster and Java TreeView to visualize results.[48] To determine whether the results are robust, we used three different cutoffs for standard deviation resulting in three different lists of CpG sites to cluster: (A) 0.01 (18,815 sites), (B) 0.10 (11,205 sites), and (C) 0.40 (2,457 sites). The range of standard deviations was 0.00017 - 0.57. Complete linkage and single linkage resulted in the same conclusions.

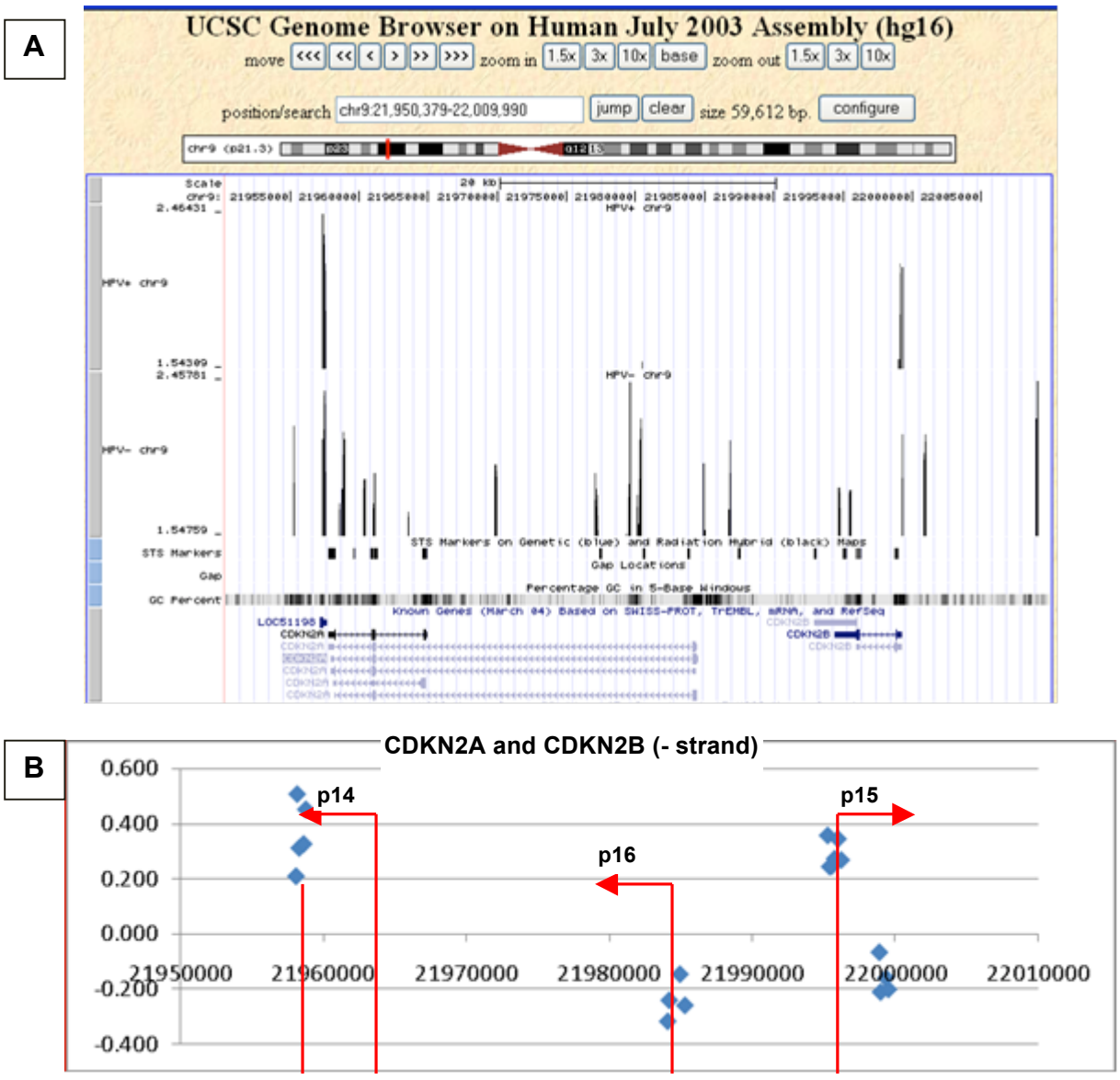


Figure S3: (A) UCSC Genome Browser view of *Cdkn2a* and *Cdkn2b* for averaged HPV(+) cells (top track) and averaged HPV(-) cells (middle track). Figure shows overall higher methylation in the HPV(-) cell lines. Also shown are GC percent and knownGene transcripts (bottom 2 tracks) **(B)** Schematic of the Illumina Infinium HumanMethylation27 results for for *Cdkn2a* (*p16*) and *Cdkn2b* (*p15*). Blue diamonds indicate difference in % methylation for HPV(+) – HPV(-) cell lines at sites measured by the Illumina platform. Vertical red lines indicate locations of transcription start sites and direction of transcription.

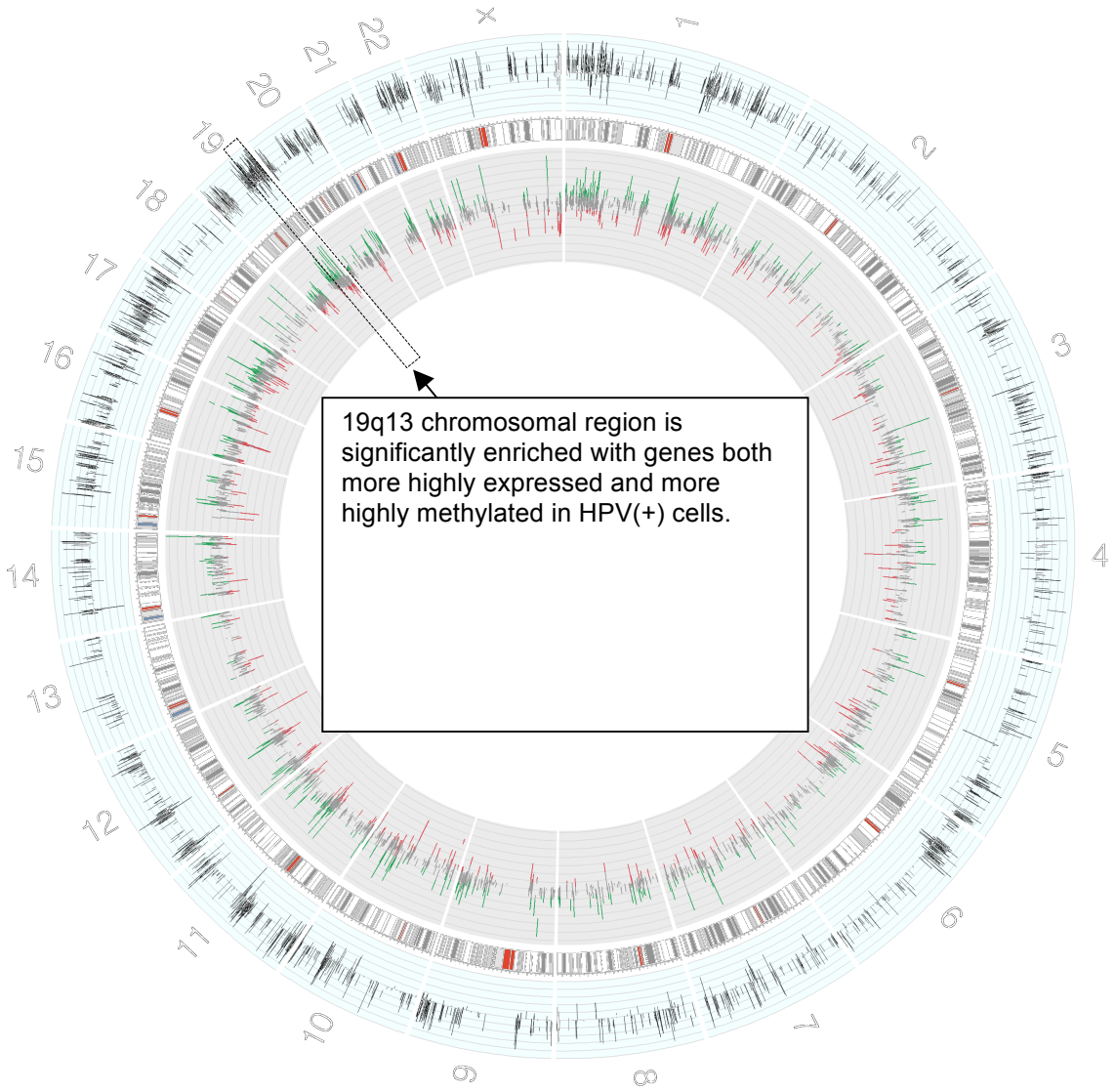
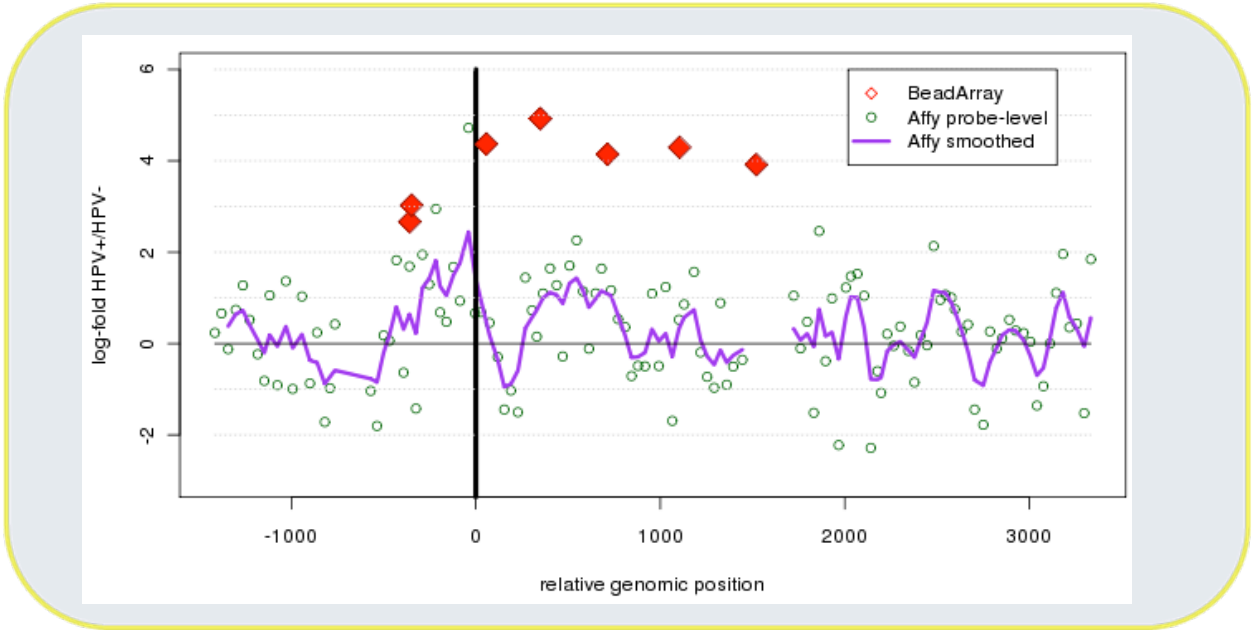


Figure S4. Circos plot of genome-wide DNA methylation and gene expression differences between HPV(+) and HPV(-) cell lines. The chromosome numbers are given on the periphery of the plot. The outer ring (blue background) depicts the average difference in DNA methylation between HPV positive and negative cell lines. The middle ring depicts the chromosome bands of the p and q arms with the centromere marked in red. The inner ring (grey background) depicts average differences in gene expression, with sites with at least an average two fold increase in expression in HPV(+) colored green and those with at least a two fold decrease in expression in HPV(+) colored red.



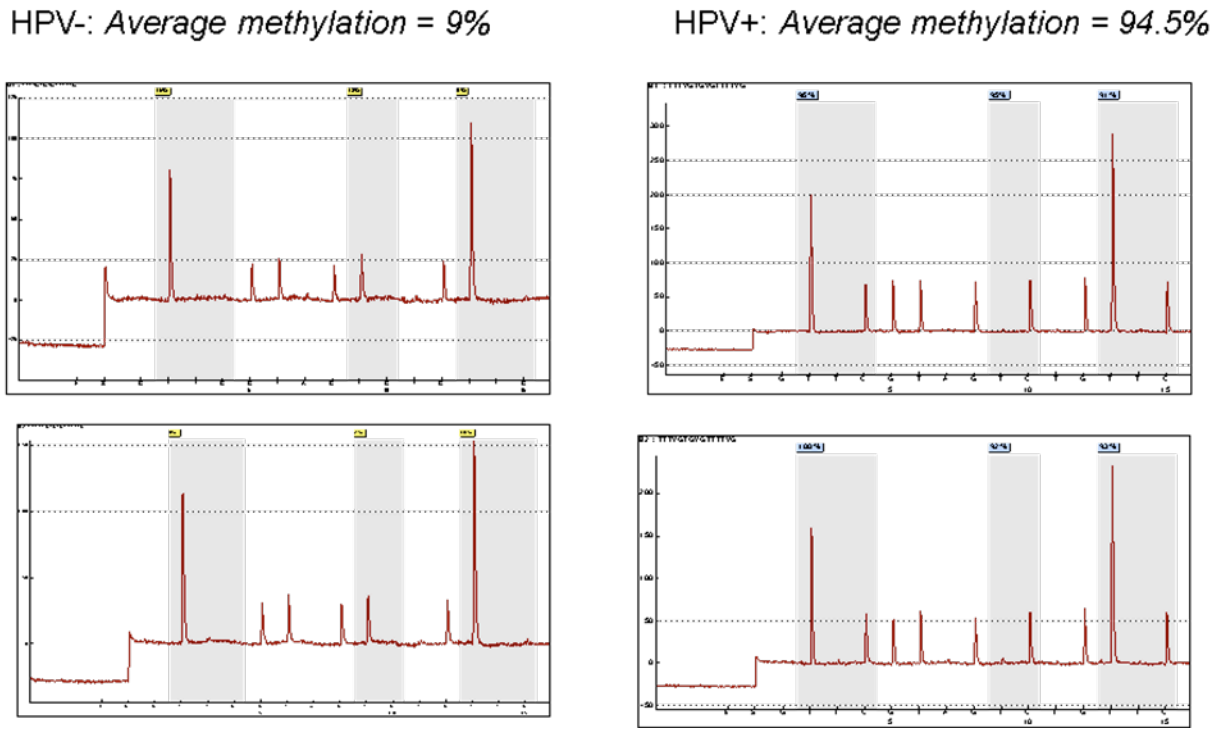


Figure S5: (A) Illumina BeadArray estimated *ESR1* methylation of 4% in HPV(-) tumors compared to 82% in HPV(+) tumors closest to TSS. This was in good agreement with the Affymetrix Chip Set, which estimated a 25-fold increase in methylation in the HPV(+) tumors at the most significant probe , just upstream of the start site**. (B)** Pyrosequencing validates *ESR1* as differentially methylated in HPV+ and HPV- HNSCC in the original four cell lines.

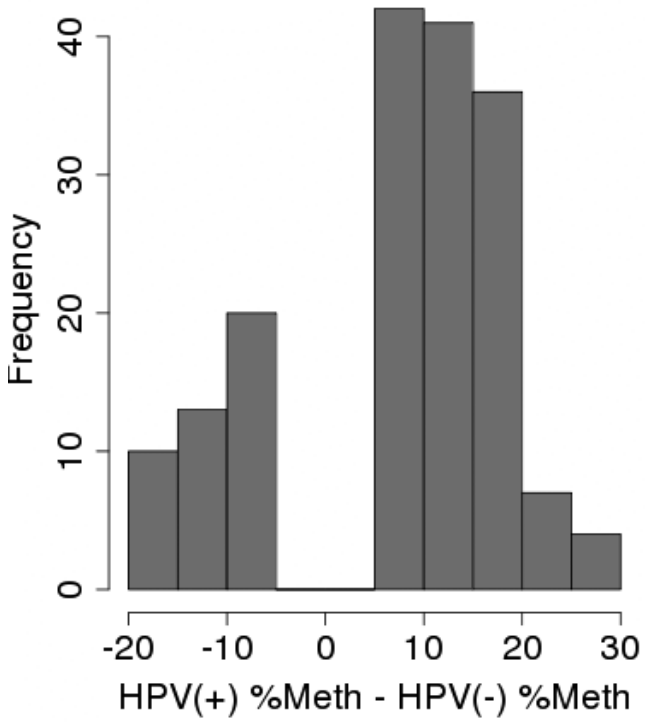
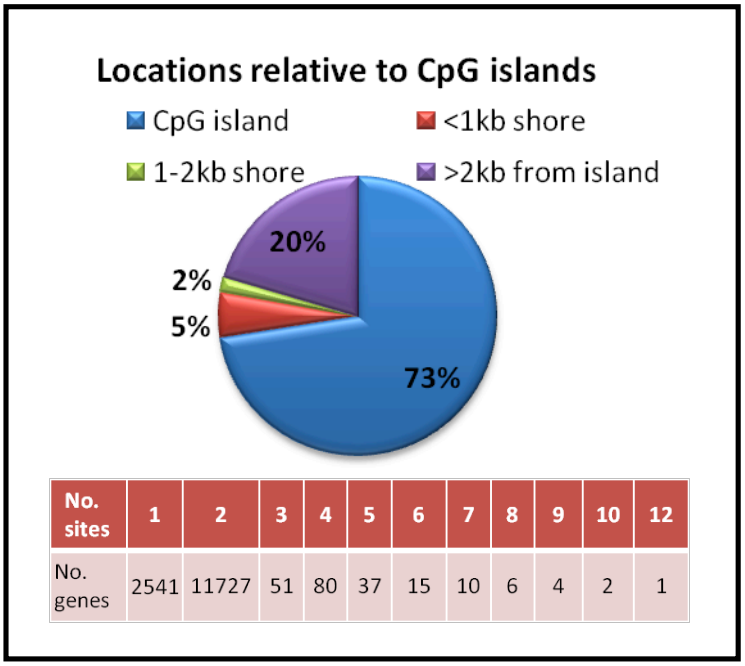
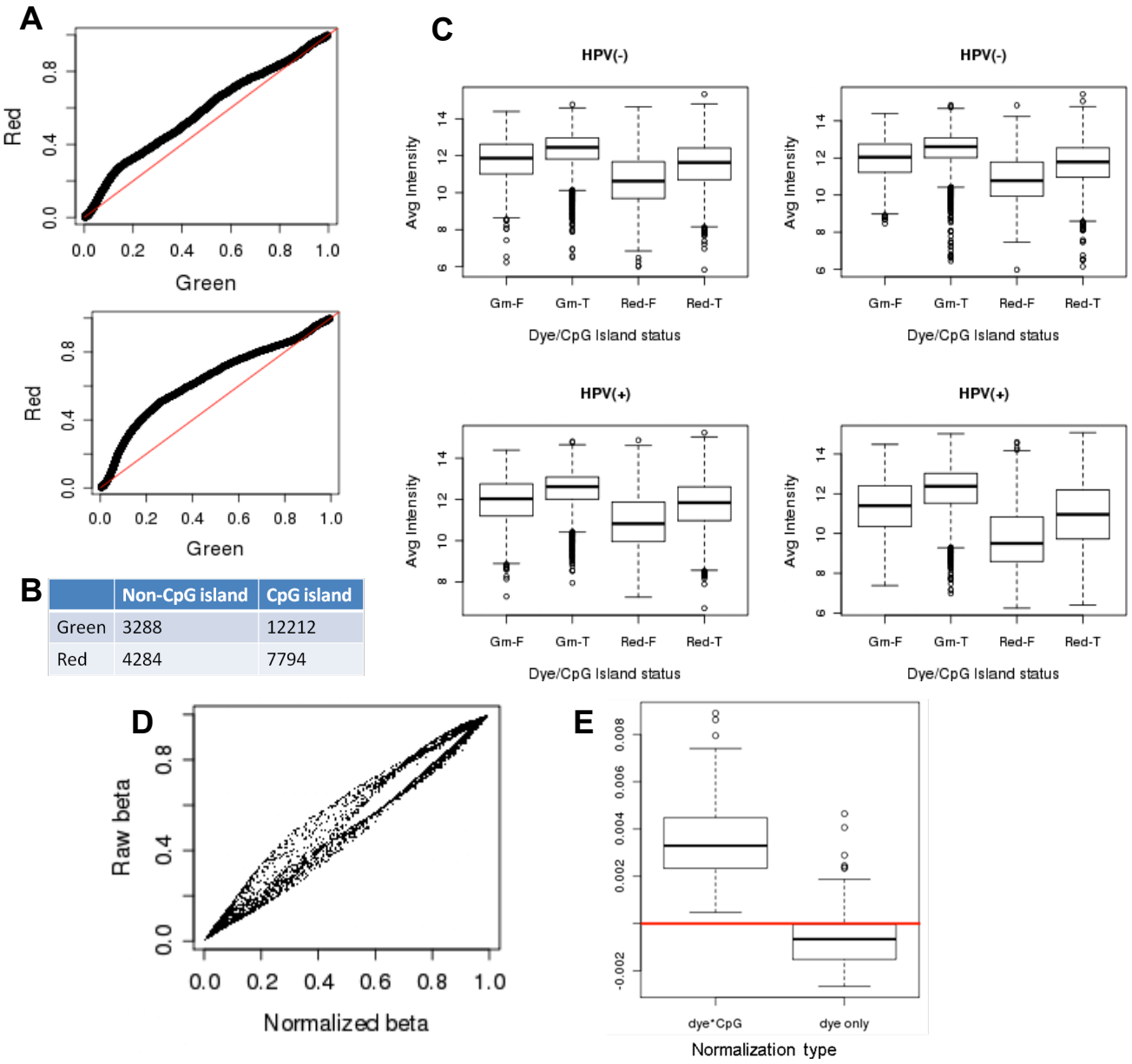


Figure S6: Histogram of differences in % methylation between HPV(+) (n=18) and HPV(-) (n=28) tumor samples shows a similar trend towards higher methylation in HPV(+) samples as was observed in cell lines. CpG methylation was measured using the Illumina Goldengate Cancer Panel. CpG sites were filtered by p-value < 0.10 for having a different % methylation in HPV(+) versus HPV(-) tumor samples, analyzed by t-tests.



No. sites	1	2	3	4	5	6	7	8	9	10	12
No. genes	2541	11727	51	80	37	15	10	6	4	2	1

Figure S7: Properties of the Illumina HumanMethylation27 Infinium BeadArray. Top: Pie chart showing the relative proportions of probes in CpG islands, within 1kb of a CpG island, 1-2kb from a CpG island, and farther than 2kb. **Bottom:** Table showing that most genes are represented by 1 or 2 sites (probes) on the array. A small percent, mainly imprinted and cancer-related genes, are represented by up to one dozen sites (probes).



	Non-CpG island	CpG island
Green	3288	12212
Red	4284	7794

Figure S8: Motivation and results for our novel dye normalization method for the Illumina HumanMethylation27 Infinium BeadArray. (A) Example QQ-plots of the red versus green beta values (estimates of % methylation) show a non-linear relationship with green dye stronger in the low expression range **(B)** Table showing the counts of CpG island and non-CpG island probes labeled in red (Cy5) and green (Cy3). Numbers demonstrate that dye is confounded with CpG island status **(C)** Boxplots of average log intensities, divided by dye and CpG island status (T=in CpG island; F=not in CpG island) for each of the four cells lines **(D)** Scatter plot of the normalized versus raw beta values. As the plot illustrates, the largest differences are in the mid-range of % methylation. **(E)** Boxplots showing improvement in correlation between 18 replicate sample pairs from a separate study of colon tumors after our normalization scheme (left), however, no improvement in correlations is observed when performing a similar quantile normalization without taking CpG island status into account (right). There was a small, but significant, improvement in p-values such that the average improvement was 0.004, or ~1.0% ($p=3.0 \times 10^{-9}$).

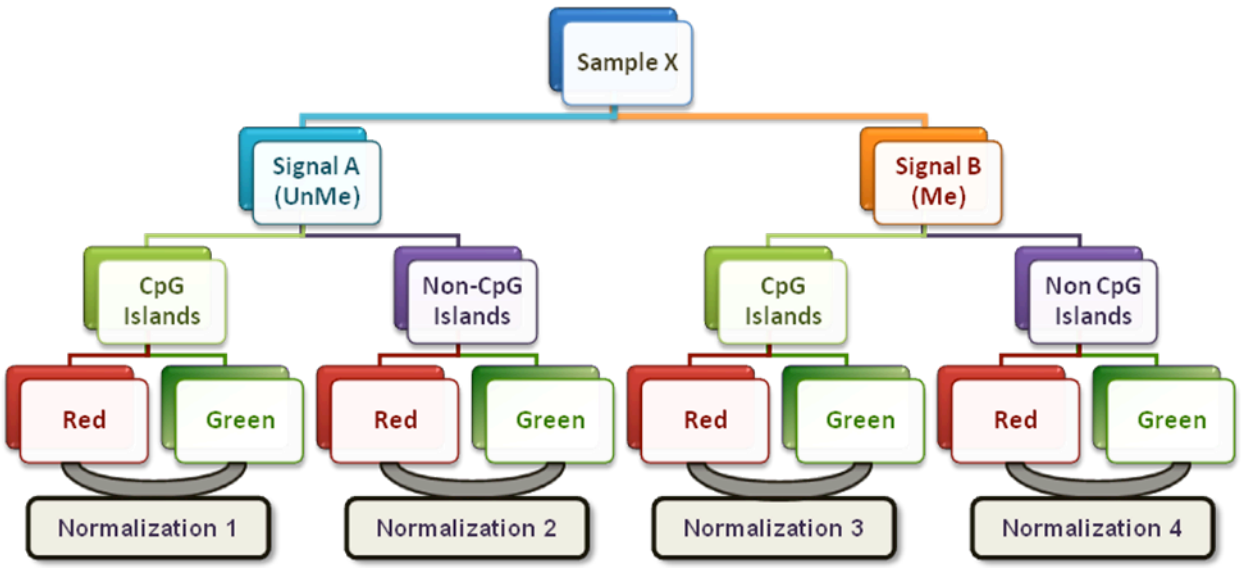


Figure S9: Strategy for the dye normalization method for the Illumina HumanMethylation27 Infinium BeadArrays. The Unmethylated (Signal A) and Methylated (Signal B) probes for each sample are divided into groups according to whether they are in CpG islands or not, and whether they are labeled with Red or Green dye. The red and green signal intensities are then quantile normalized, allowing the overall distribution of % methylation to remain the same for each sample, and within CpG island status.

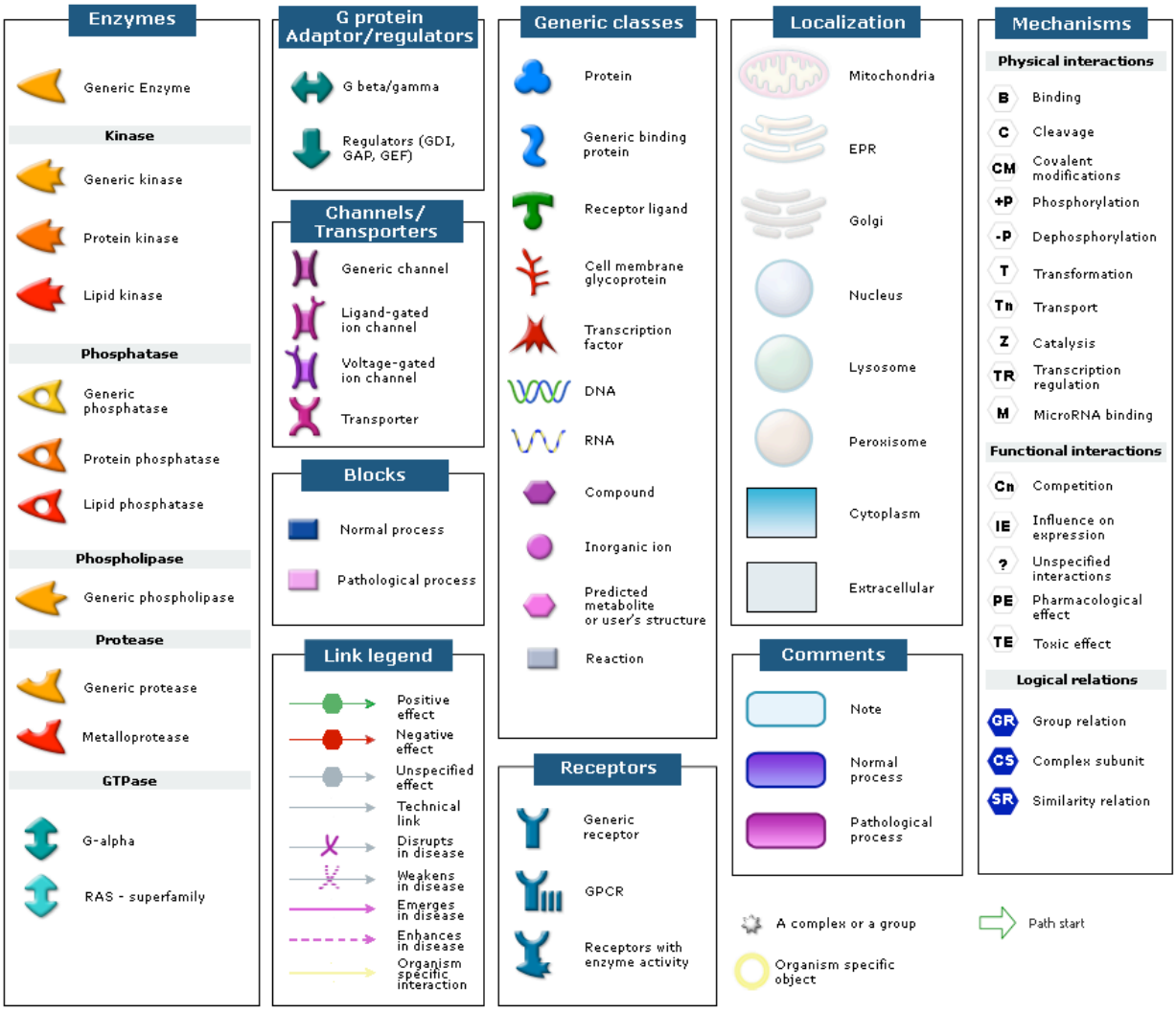


Figure S10: GeneGO MetaCore legend for use with Figures 3 and 4.

Supplemental Tables

Supplementary Table 1: Primers for pyrosequencing validation and LINE-1. The *DCC* PCR amplicon is 155 base pairs and contains 5 CpG sites. The *ESR1* PCR amplicon is 114 base pairs and contains 6 CpG sites; we quantified methylation at 3 CpG sites. PCR was carried out using HotStar Taq Master Mix (Qiagen, Valencia, CA) with 3µL of bisulfite converted DNA for 45 cycles to ensure exhaustion of biotinylated primers.

Gene	Forward primer	Reverse primer	Sequencing primer
DCC	5'-GGTTGGTTGATTAGGATTGTTGTAATT	5'-biotin-CCCCTCACTATACCCCAATACCCATCTA	5'-TTGATTAGGATTGTTGTAATTT
ESR1	5'-AGTTTTTTTTGGGTTATTTTTAGTAGAT	5'-biotin-AAACAACTTCCCTAAACTTTACTTTAC	5'-TTTTTGGGTTATTTTTAGTAGAT
LINE-1	5'-TTGAGTTAGGTGTGGGATATAGTT	5'-CAAAAAATCAAAAAATTCCCTTTCC	5'- AGGTGTGGGATATAGT

Supplementary Table 2. Selected differentially expressed and CpG methylated genes between HPV+ and HPV- cell lines

Entrez ID	Gene symbol	Chromosomal location	% meth in HPV+	% meth in HPV-	% change in methylation	Methylation p-value	Expression Fold change	Expression p-value
Higher Methylation / Lower Expression in HPV-								
3856	KRT8	Chr12:51585412	13%	80%	67%	0.0018	13.5	0.031
3204	HOXA7	Chr7:27162678	6%	93%	87%	0.00062	5.6	0.0057
1029	CDKN2A	Chr9:21984108	2%	34%	32%	0.013	48.5	0.0012
5446	EHF	Chr11:34599461	7%	45%	38%	3.53E-05	48.6	0.0039
29984	RHOD	Chr11:66581226	5%	71%	66%	0.022	5.1	0.078
8835	SOCS2	Chr12:92490616	4%	52%	48%	0.042	4.4	0.090
8416	ANXA9	Chr1:149220567	31%	81%	50%	0.017	11.6	0.00014
283212	KLHL35	Chr11:74818329	1%	97%	96%	5.14E-06	7.1	5.1E-06
27141	CIDEB	Chr14:23850391	7%	87%	79%	0.0076	6.2	0.015
84842	HPDL	Chr1:45565129	2%	80%	78%	0.025	14.8	0.00060
1902	LPAR1/EDG	Chr9:112841141	1%	80%	79%	0.0080	8.2	0.0025
4291	MLF1	Chr3:159771755	2%	67%	66%	0.0070	3.8	0.045
9743	RICS	Chr11:12839906	15%	76%	60%	0.046	2.5	0.035
7980	TFPI2	Chr7:93358381	29%	89%	61%	0.013	53.8	0.0011
Higher Methylation / Lower Expression in HPV+								
8900	CCNA1	Chr13:35904611	59%	3%	56%	0.0017	-7.1	0.040
230	ALDOC	Chr17:23928055	80%	5%	75%	0.0090	-4.3	0.0048
1290	COL5A2	Chr2:189752881	93%	5%	88%	5.43E-05	-48.1	0.0031
1514	CTSL1	Chr9:89530472	91%	7%	84%	0.00097	-4.1	0.046
2517	FUCA1	Chr1:24067730	83%	31%	51%	0.030	-4.7	0.0060
3257	HPS1	Chr10:10019691	93%	70%	23%	0.0046	-2.5	0.0213
3667	IRS1	Chr2:227374043	96%	48%	48%	0.0091	-2.2	0.031
4000	LMNA	Chr1:154349900	83%	26%	58%	0.00039	-1.9	0.035
126308	MOBKL2A	Chr19:2046364	97%	7%	90%	0.00055	-2.6	0.034
84545	MRPL43	Chr10:10273853	91%	66%	26%	0.00049	-3.2	0.0079
26471	NUPR1/P8	Chr16:28457672	91%	11%	80%	3.50E-05	-19.2	0.0085
11142	PKIG	Chr20:42593663	66%	5%	61%	0.0041	-2.3	0.039
6781	STC1	Chr8:23768300	65%	11%	54%	0.0032	-145.1	2.5E-05
57415	C3orf14	Chr3:62280187	95%	4%	91%	2.61E-05	-50.7	0.00032
9976	CLEC2B	Chr12:9913949	37%	9%	28%	0.026	-9.7	0.0049
2069	EREG	Chr4:75449765	56%	4%	52%	0.00052	-42.7	0.022
1656	DDX6	Chr11:11816799	77%	17%	60%	0.025	-1.3	0.024
2799	GNS	Chr12:63440249	81%	11%	70%	0.0090	-2.9	0.0079
1734	DIO2	Chr14:79747441	97%	7%	90%	0.00014	-7.0	0.0058
3572	IL6ST	Chr5:55326503	48%	6%	42%	0.0075	-2.2	0.016
55745	MUDENG	Chr14:56804217	90%	38%	52%	0.043	-2.6	0.0086
91663	MYADM	Chr19:59061281	94%	4%	90%	1.07E-05	-7.9	0.0033
23433	RHOQ	Chr2:46623803	88%	4%	84%	0.0054	-5.6	0.026
5228	PGF	Chr14:74492357	73%	12%	61%	0.036	-2.2	0.020
5229	PGGT1B	Chr5:114626745	82%	22%	60%	0.031	-1.4	0.028
5999	RGS4	Chr1:161305779	80%	5%	75%	0.015	-52.5	0.026
10556	RPP30	Chr10:92621408	75%	12%	62%	0.00084	-2.2	0.0039
7170	TPM3	Chr1:152431575	82%	22%	61%	0.013	-2.7	0.017

Chromosomal location is given in human genome version hg17 coordinates.

Supplementary Table 3. Top ranked concepts significantly enriched with genes up- or down- regulated in HPV+ compared to HPV- cell lines

Concept Type	Concept Name	# genes in concept	Odds ratio	p-value	FDR	Genes with p<0.05 for expression difference
Up in HPV+ cells						
Cytoband	9p24.1	16	129.0	6.3E-17	3.7E-14	UHRF2, C9ord123, C9orf46, CDC37L1, RLN2, KIAA1432, PPAPDC2
Cytoband	19q13.11	28	32.8	3.0E-10	4.7E-08	NUDT19, RHPN2, CEBPG, PDCD2L
	19q13.12	55	16.5	1.3E-09	1.1E-07	ZNF260, HAUS5, ZNF571, ZNF383, ZNF567, ZNF566
MiMI	SFN interactions	124	9.3	6.3E-10	2.8E-06	PTPN3, CGN, RHPN2, KRT18
MeSH	Desmosomes	28	29.9	1.6E-09	3.4E-06	KLK5, DSG4, KRT18, KRT8
MeSH	Keratinocytes	72	11.9	9.3E-09	6.8E-06	FLG, SCEL, CRABP2, IVL, IL1A
GO	Epidermis development	150	6.7	1.4E-08	2.1E-05	FST, HOXA7, FLG, PTGS2, SCEL, CYP27B1, CST6, KLK5, NGFR, IVL, CRABP2, CDKN2A, WNT7A, LAMC2
MiMI	SELL interactions	16	44.7	5.1E-08	7.6E-05	SELE, PRKCQ, PRKCI, MUC7
MeSH	Epithelial cells	60	11.9	1.4E-07	5.9E-05	PTGS2, OCLN, CCL28, LAMC2, KRT18
Up in HPV- cells						
Cytoband	16p11.2	101	18.5	1.6E-12	4.7E-10	PRSS53, VKORC1, NUPR1, C16orf54, ALDOA
Cytoband	Xq28	93	15.7	3.2E-10	4.7E-08	GABRE, GABRA3, G6PD, MPP1
Pfam	Core histone H2A/H2B/H3/H4	39	33.0	1.3E-09	4.6E-07	HIST1H4F, HIST2H2BE, HIST1H2BM
KEGG	Systemic lupus erythematosus	103	14.1	8.1E-09	1.5E-06	HLA-DRA, HIST1H4F, HIST2H2BE, HIST1H2BM, HIST2H2AA3, C1S
GO	Calcium-dependent cell-cell adhesion	24	55.5	7.8E-10	2.8E-06	CDH2, PCDHB2, PCDHB10,
MeSH	3-Hydroxysteriod Dehydrogenases	12	98.2	6.3E-09	6.8E-06	HSD3B2, AKR1C3, AKR1C2, AKR1C1
GO	Fibrillar collagen	11	100.2	1.7E-08	2.1E-05	COL5A2, LUM
GO	Calcium ion binding	861	2.6	4.6E-08	4.2E-05	CACNA2D3, HPCAL1, VWCE, GNS, PLCB1, EMR1, DTNA, DSPP, NELL2, CDH18, CDH2, C1S, BGLAP, MYL6B, ITPR1, PCDHB2, PCDHB8, PCDHB10, CGREF1,

Enrichment tested with LRpath . MiMI = *Michigan Molecular Interactions* comprehensive database of protein interactions. MeSH-defined gene groups were defined based on http://gene2mesh.ncibi.org.

Supplemental Methods

DNA Methylation with Infinium HumanMethylation27 BeadArray

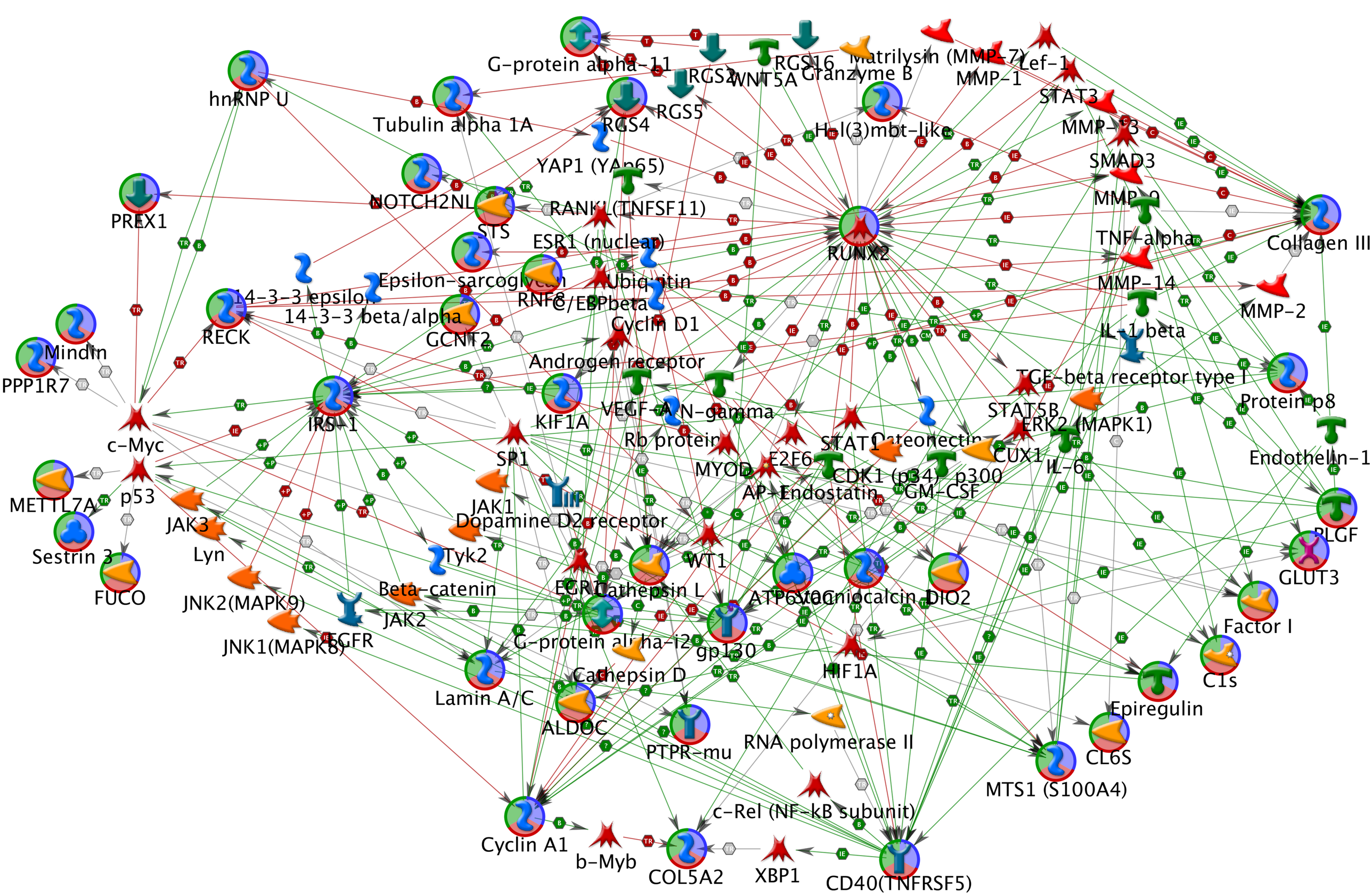
The Illumina Infinium HumanMethylation27 BeadArray platform is inherently different from other Cy3/Cy5 platforms. Whereas for most Cy3/Cy5 platforms, the measurements being compared are labeled with the two different dyes, in this platform the measurements being compared (methylated versus unmethylated DNA from the same site) are labeled with the same dye. In this platform, the dye is instead determined by the nucleotide base following the CG site: red for T or A and green otherwise. Thus, it may seem reasonable to assume that no dye normalization is needed for this platform. However, as we explain below, our observations of the properties of this data and additional data from this platform, led us to conclude otherwise. However, since the two measurements directly being compared on this platform are labeled with the same dye, we cannot perform the usual local regression or smoothing spline to normalize for dye effect. Following is the motivation for performing a dye normalization for this platform. Although the sets of red and green-dye labeled probes could be thought of as two separate experiments, their beta values (estimating % methylation) have two different, likely non-linear, relationships with actual % methylation. This non-linear relationship is hinted at by the quantile-quantile plots (qq-plots) for raw % methylation between red and green sites presented in Supplemental Figure 2A. Similar to what is often observed in other Cy3/Cy5 platforms, Cy3 (green) is observed as being stronger in the low expression range. However, one may argue that since dye is determined by the nucleotide base following the CG pair, this difference may have a meaningful biological basis rather than being a technical factor, and therefore should not be corrected for. If the following base is a C or G, the region may be more likely to be a CpG island. Indeed, the ratio of CpG island:non-CpG island for green is 3.71, whereas the ratio of CpG island:non-CpG island for red is 1.82, meaning that a green labeled probe is twice as likely to be in a CpG island than a red labeled probe (Supplemental Figure 2B). Thus, confounding between dye and CpG island status exists. However, as shown in Supplementary Figure 2C, the average intensity of methylated and unmethylated DNA is affected *both* by dye and CpG island status, with CpG islands and green-labeled sites tending to have higher fluorescent levels. Since the dye-affect still exists, even after accounting for CpG island status, this will affect the estimated % methylation and differences in % methylation between two samples, and thus will affect the ranking of the results. Since many investigators concentrate on ranks of genes, whether by % methylation, change in % methylation or significance, we should normalize the dyes to have the same relationship with actual % methylation, even if this true relationship is unknown. Otherwise, the red and green labeled CG sites will display different distributions for change in % methylation, skewing the ranked results.

Our normalization method was developed to correct for the non-linear relationship between beta values from red versus green-labeled sites, while NOT affecting the overall distribution of % methylation in CpG island sites and non-CpG island sites for each sample. Our normalization consists of the following steps, and was implemented using R statistical environment and Bioconductor. First, for each sample, obtain the raw signal levels for methylated (unbisulfite converted) and unmethylated (bisulfate converted) probes that are in a CpG island separately for green and red-labeled probes. Second, perform quantile normalization for the red versus green CpG island probes for each sample separately. Next, repeat this type of quantile normalizations for the non-CpG island probes. Since the number of red and green-labeled

probes differ, “NA” values were added to the set containing fewer probes to allow the use of the normalize.quantiles function in R/Bioconductor software. The signal levels for CpG and non-CpG islands are then recombined, and the normalized beta values are defined as B/(A+B), where B = the quantile normalized intensity for methylated DNA, and A = the quantile normalized intensity for unmethylated DNA. A scatter plot of raw versus normalized beta values is shown in Supplemental Figure 2D. To test whether this normalization method has a benefit, we used it to normalize 18 additional, unrelated colon cancer samples from this platform, which allowed us to obtain a better estimate of improvement in correlation between replicate samples. Our novel normalization procedure not only improved the correlation among replicate samples in this experiment and the larger set of colon cancer samples, but also slightly improved the p-values for differential methylation.

Gene Expression Analysis

As stated, functional enrichment testing of the expression data was performed using LRpath[21] . LRpath is a logistic regression-based method shown to perform favorably compared to alternative approaches in small sample experiments[21]. Because the original version of LRpath is limited to testing Gene Ontology (GO) and KEGG pathways, we developed a custom implementation of LRpath that incorporates multiple additional gene groups, including cytoband locations and several obtained from the database for the web-based gene set enrichment program, ConceptGen[22], in additional to GO and KEGG. The additional gene groups are the MiMI protein interactions, Pfam, MeSH, Biocarta, and Panther concept types from ConceptGen.


G-protein alpha-11
RGS2
RGS16
WNT5A
Granzyme B
Matrilysin (MMP-7)
Lef-1
MMP-1
STAT3
hnRNP U
Tubulin alpha 1A
RGS4
RGS5
H-l(3)mbt-like
MMP-13
SMAD3
YAP1 (YAp65)
NOTCH2NL
RANKL(TNFSF11)
MMP-9
PREX1
STS
ESR1 (nuclear)
RUNX2
TNF-alpha
Collagen III
Epsilon-sarcoglycan
Ubiquitin
MMP-14
14-3-3 epsilon
RNF8
C/EBPbeta
MMP-2
14-3-3 beta/alpha
RECK
GCNT2
Cyclin D1
IL-1 beta
Mindin
Androgen receptor
PPP1R7
TGF-beta receptor type I
IRS-1
KIF1A
VEGF-A
IFN-gamma
STAT5B
Protein p8
ERK2 (MAPK1)
Rb protein
STAT1
Osteonectin
SP1
E2F6
CUX1
IL-6
Endothelin-1
MYOD
CDK1 (p34)
p300
METTL7A
p53
AP-Endostatin
GM-CSF
JAK1
Dopamine D2 receptor
PLGF
JAK3
Lyn
Tyk2
WT1
Sestrin 3
ATP6V0C
Osteocalcin
DIO2
GLUT3
FUCO
Beta-catenin
EGR1
Cathepsin L
Factor I
JNK2(MAPK9)
JAK2
gp130
G-protein alpha-i2
JNK1(MAPK8)
EGFR
HIF1A
C1s
Cathepsin D
Lamin A/C
Epiregulin
ALDOC
PTPR-mu
RNA polymerase II
CL6S
MTS1 (S100A4)
c-Rel (NF-kB subunit)
Cyclin A1
b-Myb
COL5A2
XBP1
CD40(TNFRSF5)