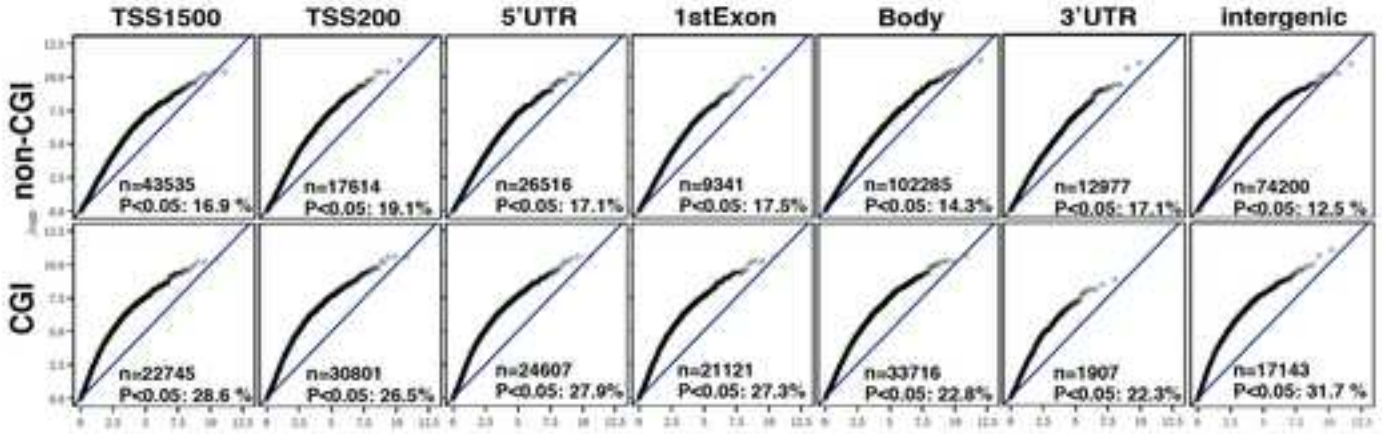
supplement figure1
Click here to download high resolution image

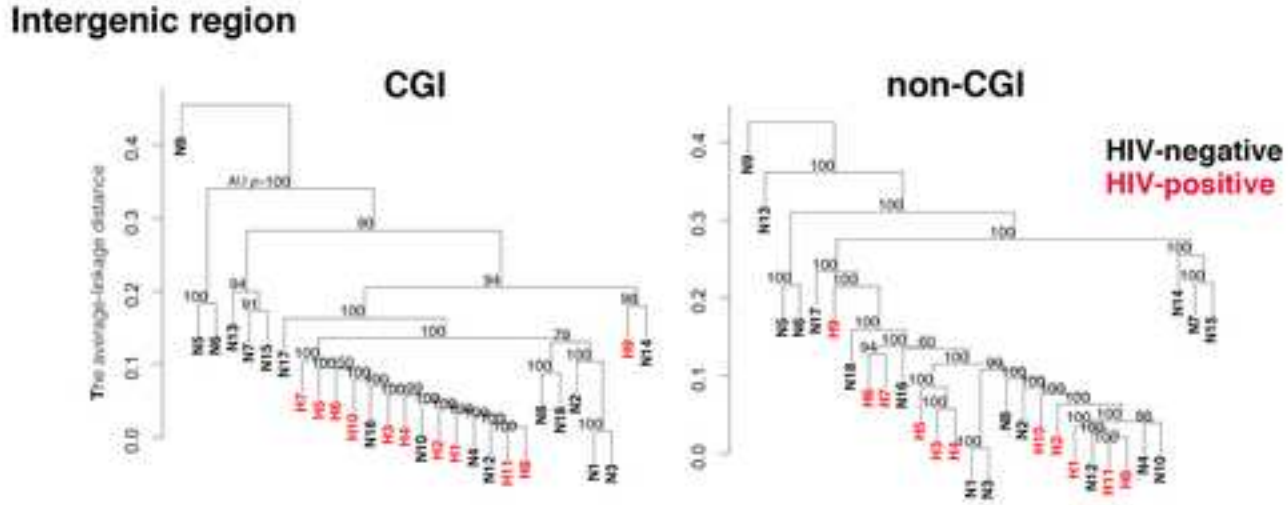
Supplementary Fig. 1



Supplementary Fig. 1. Quantile-Quantile (Q-Q) plots of CpG island (CGI) and non-CGI methylation in various gene and intergenic regions comparing HIV-associated and non-HIV lymphomas in Cohort I. The oblique line in each panel indicates the null hypothesis that there is no significant difference in methylation between HIV-associated and non-HIV lymphomas. n, number of examined probes; $p < 0.05$ (%), proportion of significant probes.

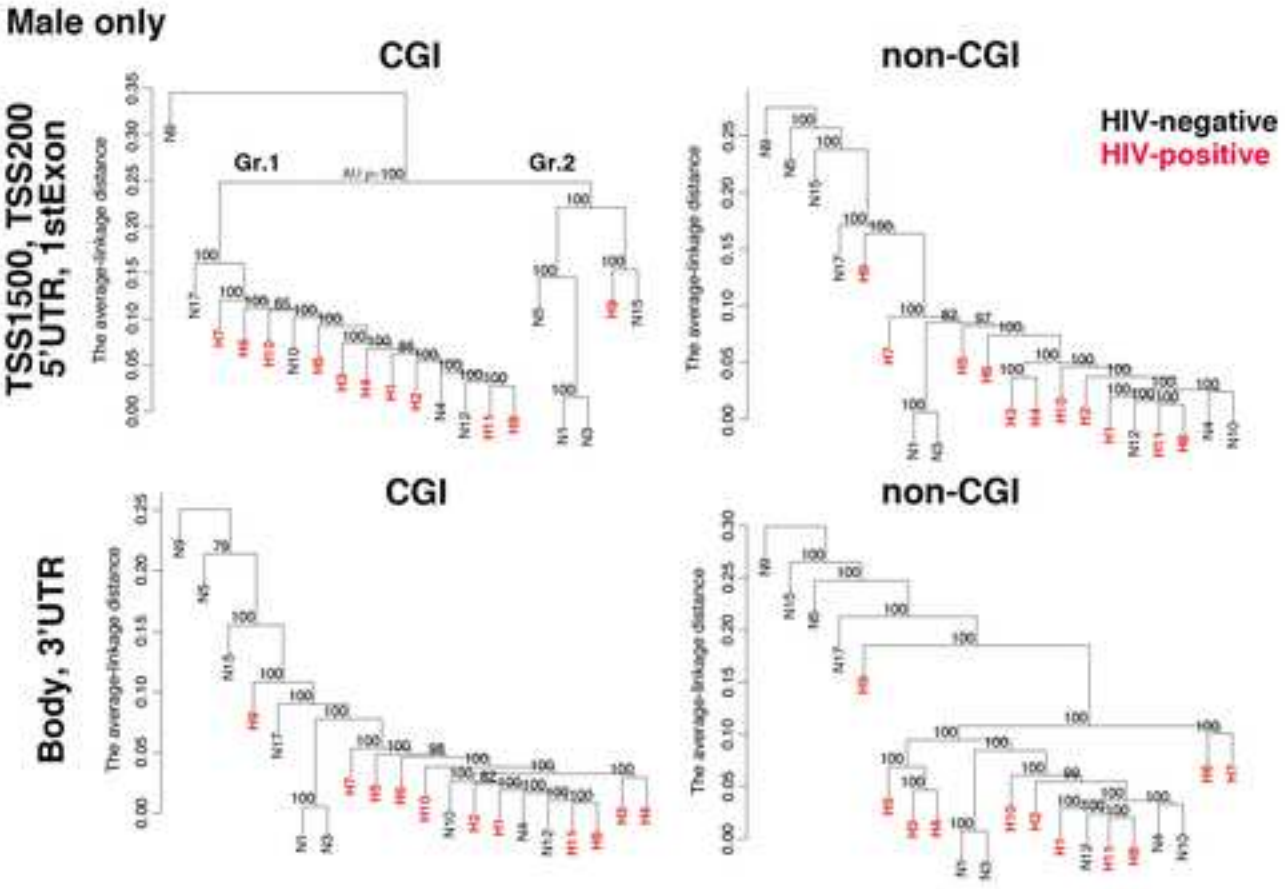
supplement figure2
Click here to download high resolution image

Supplementary Fig. 2



Supplementary Fig. 2. Methylation profile analysis of HIV-associated and non-HIV lymphoma DNA by intergenic region in Cohort I using Infinium HumanMethylation450 BeadChip technology. Hierarchical clustering analysis of CpG island (CGI) and non-CGI methylation in the intergenic region of lymphoma DNA. AU *p*-value, approximately unbiased *p*-value computed using multi-scale bootstrap resampling.

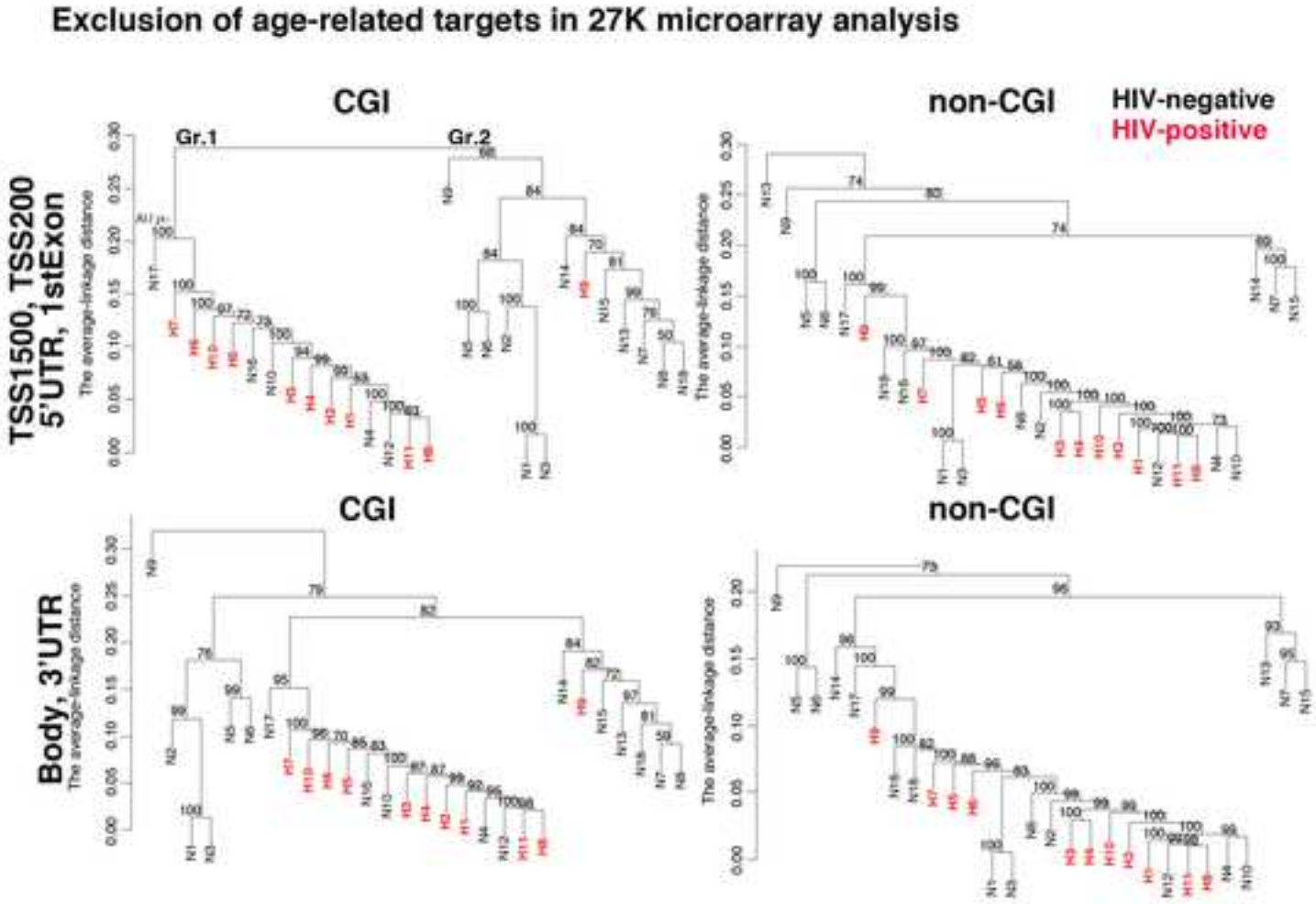
supplement figure3
Click here to download high resolution image



Supplementary Fig. 3. Methylation profile analysis of HIV-associated and non-HIV lymphoma DNA in males only of Cohort I using Infinium HumanMethylation450 BeadChip technology. Hierarchical clustering analysis of CpG island (CGI) and non-CGI methylation of lymphoma DNA from males only. The analysis of CGI methylation produced two groups that distinguished between HIV-associated lymphomas (Group 1, Gr. 1) and non-HIV lymphomas (Group 2, Gr. 2). AU *p*-value, approximately unbiased *p*-value computed using multi-scale bootstrap resampling.

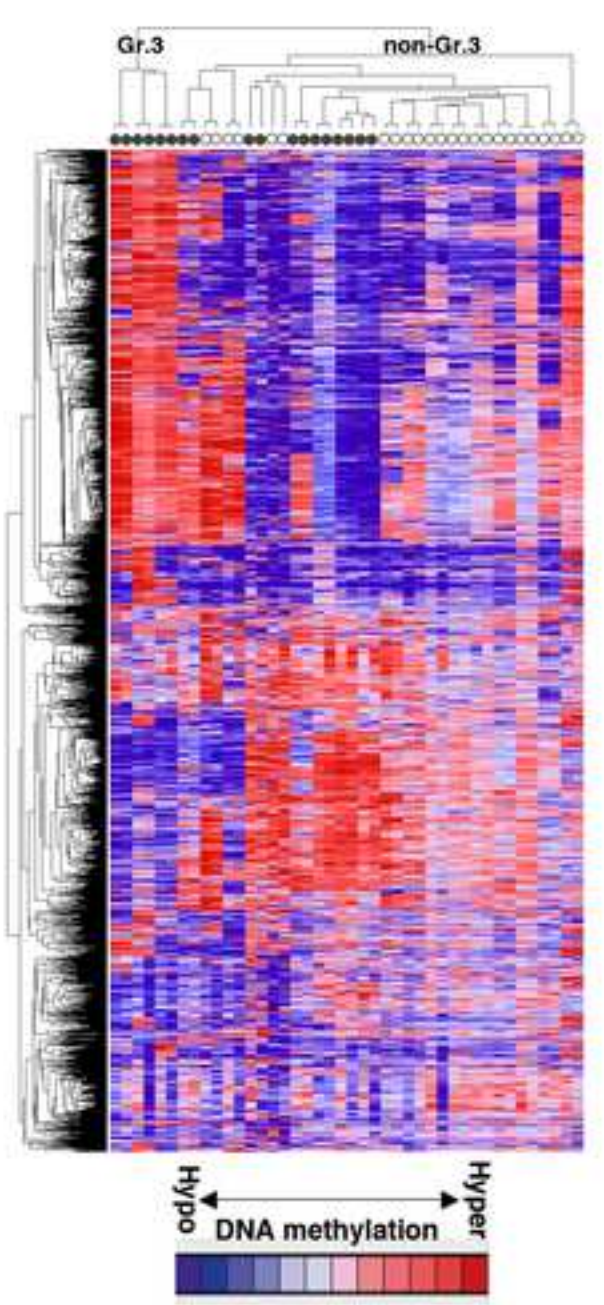
supplement figure4
Click here to download high resolution image

Supplementary Fig. 4



Supplementary Fig. 4. Methylation profile analysis of HIV-associated and non-HIV lymphoma DNA, excluding age-related targets, in Cohort I, using Infinium HumanMethylation450 BeadChip technology. Hierarchical clustering analysis of CpG island (CGI) and non-CGI methylation of lymphoma DNA excluding age-related targets, based on previous findings. The results are similar to those including the age-related targets (Fig. 1). AU *p*-value, approximately unbiased *p*-value computed using multi-scale bootstrap resampling.

supplement figure5
Click here to download high resolution image



Supplementary Fig. 5. Methylation profile analysis of recurrent HIV-associated lymphoma and non-HIV lymphoma DNA in Cohort II, using Cancer Panel I. Cancer Panel I microarray analysis was performed for nine HIV-associated lymphomas and 12 non-HIV lymphomas in Cohort II (see Supplementary Table 1). The color bar indicates hyper- and hypomethylation. Clustering analysis of the methylation profiles produced two groups: Group 3 (Gr. 3; recurrent cases), as in Figure 2, and non-Group 3. Solid circles, HIV-associated lymphoma; open circles, non-HIV lymphoma.

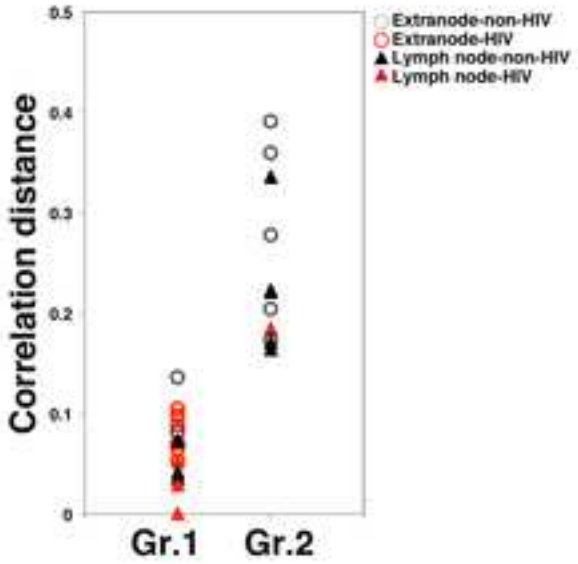
supplement figure6
Click here to download high resolution image

Supplementary Fig. 6

(a)

Tumor location*	Gr.1** (n=15)	Gr.2** (n=12)	*p*-value*** (Gr.1 vs Gr. 2)
Lymph node	6	7	
Extra-node	9	5	0.45
Stomach	0	1	
Colon	1	0	
Chest wall	1	0	
Ileocecal junction	0	2	
Skin	2	1	
Pyramid of medulla	0	1	
Brain	2	0	
Lung	1	0	
Liver	2	0	

(b)



Supplementary Fig. 6. Tumor location and sample correlation distances from methylation profile analysis in Cohort I, using Infinium HumanMethylation450 BeadChip technology. (a) Frequency of lymph node tumors, extra-nodes, and extra-node details for Human Methylation450 in Groups 1 (Gr. 1) and 2 (Gr. 2) are shown. * Tumor location: tumor site used for the Human Methylation450 microarray analyses. ** Gr. 1 and Gr. 2: See hierarchical clustering analysis in Fig. 1b. One HIV-negative lymphoma was outside its category, and another three HIV-negative lymphomas were excluded from the analyses via the filtering steps (see Methods). ***The statistical values of differences in the categorical variables were calculated using Fisher's exact test. **(b)** Correlation distance between the case H8 in Gr.1 and other cases, which was calculated in the hierarchical clustering analysis process using TSS1500, TSS200, 5'UTR, 1stExon, and CGI methylation profiles (Fig. 1b).

suppleTable1
Click here to download high resolution image

Supplementary Table 1

Patient characteristic of lymphoma samples for Cancer Panel I

Items examined		HIV		non-HIV
		Gr. 3	Gr. 4	
Gender	Female	1	0	4
	Male	2	6	8
Age	Mean	36.66	35.00	64.83
	SD*	5.77	13.78	18.00
Histology	BL	2	3	1
	DLBCL	1	1	8
	HD	0	2	2
	FL	0	0	1
Bcl-2	+	0	1	9
	-	3	5	2
	ND**	0	0	1
Stage	I&II	0	2	3
	III&IV	3	4	9
EBV	+	1	3	3
	-	2	3	0
	ND**	0	0	9
Recurrence	+	2	0	2
	-	1†	6	10
IPI* **	0 or 1	0	1	1
	2 or 3	1	1	3
	4 or 5	2	2	6
	ND**	0	2	2

BL: Burkitt lymphoma; DLBCL: diffuse large B cell lymphoma; HD: Hodgkin's lymphoma; FL: Follicular lymphoma; EBV: Epstein–Barr Virus

*Standard deviation

**ND, not determined

***IPI: International Prognostic Index for non-HD (Stage, LDH, Performance status, age)

† a tumor mass appeared in the cervical spinal cord about 17 months later, although recurrence was not confirmed pathologically.

suppleTable2
Click here to download high resolution image

Supplementary Table 2

DAVID* analyses of lymphoma samples for Human Methylation450 (450K) microarray analysis in Cohort I

Pathway	Count	%	*p*-value
Basal cell carcinoma	16	1.44	**1.E-07**
Pathways in cancer	41	3.70	**9.E-07**
Hedgehog signaling pathway	13	1.17	**4.E-05**
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	15	1.35	**5.E-05**
Calcium signaling pathway	24	2.16	**8.E-05**
Neuroactive ligand-receptor interaction	30	2.71	**1.E-04**
Maturity onset diabetes of the young	7	0.63	**2.E-03**
Cell adhesion molecules (CAMs)	17	1.53	**2.E-03**
Melanogenesis	14	1.26	**3.E-03**
Wnt signaling pathway	18	1.62	**3.E-03**
Adherens junction	11	0.99	**9.E-03**
Regulation of actin cytoskeleton	21	1.89	**1.E-02**
Focal adhesion	20	1.80	**1.E-02**
Tight junction	15	1.35	**1.E-02**
ECM-receptor interaction	11	0.99	**2.E-02**
Gap junction	11	0.99	**2.E-02**
Axon guidance	13	1.17	5.E-02
Heparan sulfate biosynthesis	5	0.45	5.E-02
Dilated cardiomyopathy	10	0.90	6.E-02
Vascular smooth muscle contraction	11	0.99	9.E-02
Melanoma	8	0.72	9.E-02
Hypertrophic cardiomyopathy (HCM)	9	0.81	9.E-02
MAPK signaling pathway	21	1.89	9.E-02

In total, 2541 target sites were extracted using Wilcoxon's rank sum test ($p < 0.05$) and $\Delta\beta$ values >0.30 in comparisons between HIV-associated and non-HIV lymphomas. Subsequently, 2118 targets were extracted from UCSC_REFGENE_ACCSESSION for DAVID analyses.

*DAVID (v6.7), the database for annotation, visualization, and integrated discovery (http://david.abcc.ncifcrf.gov/), was used for gene-annotation enrichment analysis and biological pathway mapping. Count: number of the 2118 genes that belong to the corresponding pathway. %: the percentage of the 2118 genes in the corresponding pathway.

p-value: Fisher's exact test for the null hypothesis (H0) that the percentage of the differentially methylated 2118 target genes on the corresponding pathway is the same as the percentage of the 30,000 human genes that belong to the corresponding pathway.

Bold *p*-values indicate statistical significance ($p < 005$).

suppleTable3
Click here to download high resolution image

Supplementary Table 3

DAVID**analyses of lymphoma samples for Cancer Panel I in Cohort II

Pathway	1421 probes†		Differentially methylated 439 probes††		
	Count (x)	%	Count (y)	%	*p*-value
Pathways in cancer	120	16.42	49	16.44	1.E-01
Cytokine-cytokine receptor interaction	62	8.48	30	10.07	6.E-02
Hedgehog signaling pathway	18	2.46	13	4.36	**3.E-02**
Focal adhesion	49	6.70	24	8.05	9.E-02
Melanoma	30	4.10	14	4.70	2.E-01
Basal cell carcinoma	19	2.60	12	4.03	6.E-02
Bladder cancer	20	2.74	10	3.36	3.E-01
Hematopoietic cell lineage	26	3.56	13	4.36	2.E-01
Adherens junction	21	2.87	10	3.36	3.E-01
Glioma	19	2.60	9	3.02	3.E-01
Colorectal cancer	31	4.24	10	3.36	9.E-01
Endocytosis	25	3.42	15	5.03	6.E-02
Allograft rejection	17	2.33	6	2.01	8.E-01
Jak-STAT signaling pathway	30	4.10	13	4.36	4.E-01
ECM-receptor interaction	16	2.19	9	3.02	2.E-01
p53 signaling pathway	25	3.42	8	2.68	1.E+00
MAPK signaling pathway	56	7.66	18	6.04	9.E-01
Prostate cancer	28	3.83	9	3.02	8.E-01
Pancreatic cancer	23	3.15	8	2.68	8.E-01
Axon guidance	29	3.97	11	3.69	6.E-01
Type I diabetes mellitus	19	2.60	6	2.01	1.E+00
Asthma	12	1.64	5	1.68	6.E-01
Intestinal immune network for IgA production	15	2.05	6	2.01	6.E-01
TGF-beta signaling pathway	28	3.83	8	2.68	1.E+00
Regulation of actin cytoskeleton	39	5.34	14	4.70	6.E-01
Acute myeloid leukemia	17	2.33	6	2.01	8.E-01
Dorso-ventral axis formation	9	1.23	4	1.34	5.E-01
Melanogenesis	20	2.74	8	2.68	5.E-01
Apoptosis	24	3.28	7	2.35	1.E+00
ErbB signaling pathway	24	3.28	7	2.35	1.E+00

**DAVID (v6.7), the database for annotation, visualization and integrated discovery (http://david.abcc.ncifcrf.gov/), was used for gene-annotation enrichment analysis and biological pathway mapping.

†: 1421 probes for XY chromosomes were extracted from all 1505 probes.

††: 439 probes were extracted from all 1505 probes based on the cut-off value with a false discovery rate (FDR) <0.01 and $\Delta\beta$ values >0.30 in comparisons between Groups 3 and 4.

p-value: *p*-values of Fisher's exact test for the null hypothesis (H0) that the percentage of the genes on our list in the corresponding pathway is the same as the percentage of the genes that do not belong to the corresponding pathway and all other genes that do not belong to the corresponding pathway.

Bold *p*-values indicate significance by Fisher's exact test ($p < 0.05$). 1421, X count; 430, Y count.

Supplementary Methods

Statistical analyses

Six gene regions: within 1,500 bps of a transcription start site (TSS1500), within 200 bps of a transcription start site (TSS200), the 5′ untranslated region (5′UTR), first exon (1stExon), body, and 3′ untranslated region (3′UTR) and intergenic regions following Illumina criteria were used for further analyses. Average linkage hierarchical clustering analysis was performed using Pearson's correlation coefficient as a similarity measure. To assess the reliability of the clustering results for Human Methylation450 microarray data, we calculated the approximated unbiased *p*-value (AU*p*) via multi-scale bootstrap resampling (Suzuki R, *et al*., Bioinformatics 22:1540-42, 2006) using the R pvclust package, which is commonly used in phylogenetic analysis. We generated 1,000 bootstrap samples to estimate AU*p*. The *p*-value of the cluster, which ranges between 0 and 1, indicates how strongly the cluster is supported by the data. The clustering structure of the Cancer Panel I microarray data was visualized using the HeatmapViewer module of GenePattern (Broad Institute, MA). Statistical analyses other than cluster analysis were performed with SAS (ver. 9.2, SAS

Institute, Cary, NC). The chi square test and Wilcoxon's rank sum test were used to assess the statistical significance of differences in categorical or continuous variables, respectively, between two independent groups. Fisher's exact test was applied when one of the frequency table cell values was ≤5. Two-dimensional false discovery rates (FDRs) (Ploner A, *et al*. Bioinformatics 22:556-65, 2006) were calculated for the data of Cancer Panel I to select candidate target genes for enrichment analysis because the sample size for Cancer Panel I analysis was small (n = 9). The statistical significance level was set at $p < 0.05$ and FDR < 0.01.

Enrichment analysis of target genes

From the genes measured on the HumanMethylation450 BeadChip, 2,541 target sites were extracted using Wilcoxon's rank sum test ($p < 0.05$) and $\Delta\beta$ values >0.30 in comparisons between HIV-associated and non-HIV lymphomas. Subsequently, 2,118 targets were extracted from UCSC_REFGENE_ACCSESSION for DAVID analyses. DAVID (v6.7), the database for annotation, visualization, and integrated discovery

(http://david.abcc.ncifcrf.gov/), was used for gene-annotation enrichment analysis and biological pathway mapping (Huang DW, *et al.*, Nucleic Acids Research 37:1–13, 2009). Fisher's exact test for the null hypothesis (H0) that the percentage of the differentially methylated 2,118 target genes on the corresponding pathway is the same as the percentage of the 30,000 human genes that belong to the corresponding pathway. *p*-values indicate statistical significance ($p < 0.05$). Of the genes measured on the Cancer Panel I, 1,421 probes for XY chromosomes were extracted from all 1,505 probes. From these, 439 probes were extracted based on the cut-off value with a false discovery rate (FDR) <0.01 and $\Delta\beta$ values >0.30 in comparisons between Groups 3 and 4. *p*-values indicate significance on Fisher's exact test ($p < 0.05$).

Validation by combined bisulfite restriction analysis (COBRA), and bisulfite DNA sequences

Among the genes measured on the HumanMethylation450 BeadChip, those with an absolute difference $\Delta\beta > 0.3$ and showing a significant ($p < 0.05$) expression difference between HIV-associated and non-HIV lymphoma samples, where the genes measured on the Cancer

Panel I had an absolute difference $\Delta\beta > 0.3$, and that showed an expression difference with a two-dimensional FDR <0.01 between the two clusters were used for validation by PCR and sequencing analyses. Genomic DNA was treated with sodium bisulfite using an EZ DNA Methylation kit (Zymo Research, Irvine, CA) and used as a template for combined bisulfite restriction analysis (COBRA) and bisulfite sequencing analysis, as previously reported (Xiong Z, *et al*., Nucleic Acids Res. 25:2532-4, 1997; Frommer M, *et al*., Proc. Nat. Acad. Sci. USA. 89:1827-31, 1992). The sequence data of at least seven cloned fragments per sample were examined using the program QUMA (http://quma.cdb.riken.jp/top/quma_main_j.html). Methylated and non-methylated DNA samples (Cat # D5014, human HCT116 DKO; Zymo Research) were used as controls.