Genome-wide DNA methylation analysis in peripheral blood cells and Epstein-Barr virus-transformed lymphoblastoid cell lines

谷口, 愛樹

https://doi.org/10.15017/1928615

出版情報：九州大学, 2017, 博士（理学）, 課程博士
バージョン：
権利関係：

Genome-wide DNA methylation analysis in peripheral blood cells and Epstein-Barr virus-transformed lymphoblastoid cell lines

Itsuki Taniguchi

Division of Genomics,

Medical Institute of Bioregulation,

Kyushu University

Contents

Abstract ---------- **ii**

General introduction ---------- **1**

Chapter 1: Methylation level measurement and methylation site selection -------- **6**

Chapter 2: Global difference between PBCs and LCLs ---------- **8**

Chapter 3: Association with CpG islands ---------- **16**

Chapter 4: Association with distance from transcription start site ---------- **18**

Chapter 5: Association with promoter type ---------- **21**

Chapter 6: The methylation level difference in age-associated methylation site **25**

Discussion ---------- **27**

Conclusion ---------- **28**

Acknowledgements ---------- **30**

Funding ---------- **30**

Abbreviations ---------- **30**

References ---------- **31**

Supplementary files ---------- **36**

Appendix ---------- **38**

Abstract

DNA methylation profiles in epidemiological studies may uncover the molecular mechanisms through which genetic and environmental factors contribute to the risks of multifactorial diseases. There are two types of commonly used DNA bioresource, peripheral blood cells (PBCs) and Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs), which are available for genetic epidemiological study. Recently, several groups showed a substantial difference in DNA methylation status between them using the relatively small size of paired samples. To confirm and extend the results, I here analyzed the methylation status of autosomes for 192 and 92 DNA samples obtained from PBCs and LCLs, respectively using the Human Methylation 450K array. After excluding SNP-associated methylation sites and low call sites, 400,240 sites were subjected to analysis using a generalized linear model with cell type, sex, and age as the independent variables. I found that the large proportion of sites showed lower methylation level in LCLs in comparison with PBCs, which is consistent with previous reports. I also performed gene ontology (GO) enrichment analysis with the genes containing the significantly methylated sites, and found that the GO terms correlated with development are enriched. This trend is seen in the genes whose expression are changed whether the cells are infected with EBV. Furthermore, I investigated the correlation between DNA methylation level and gene expression in the differentially methylated sites, and it is uncovered that there is no significant correlation. Therefore, the DNA methylation changes correlate with gene expression change indirectly, and there may be various factors in the regulation of gene expression. I also found that significantly different methylation sites tend to be located on the outside of CpG island and in the region relatively far from transcription start site. In addition, I observed that the methylation change of the sites in the low-CpG promoter region was remarkable. Finally, it was shown that correlation between chronological age and aging-associated methylation sites in *ELOVL2* and *FHL2* in LCLs was weaker than that in PBCs. In

conclusion, I found that the methylation levels of highly methylated sites of the low-CpG-density promoters in PBCs decreased in the LCLs, suggesting that the methylation sites located in low-CpG-density promoters could be sensitive to demethylation in LCLs. Despite being generated from a single cell type, LCLs may not always be a proxy for DNA from PBCs in studies of epigenome-wide analysis attempting to elucidate the role of epigenetic change in disease risks.

General introduction

Epidemiological study in multifactorial diseases

Recent years, human genome analysis technology is developed dramatically, and we have become possible to identify disease-related DNA methylation changes at the genome-wide level.

DNA methylation is one of the important epigenetic factors in the regulation of gene expression. In addition to sequence variants, it is increasingly accepted that this DNA modification may be implicated in the susceptibility of various multifactorial diseases (*1*–*3*).

Since accomplishment of human genome project and improvement of the gene analysis technology, many studies targeted the association between gene alteration and disease have reported. Specifically, a lot of genes responsible for the hereditary disorder have been identified. However, there is some disease which is not able to identify the responsible gene only by analyzing the genome sequence. Such disorder caused by a combination of genetic and environmental factors, therefore these are called multifactorial disorders. Whereas most of the genetic factors have the congenital effect to disorder, the environmental factors give acquired change to the genome. The genome modification from environmental factors is called epigenome.

Epigenome

The epigenome is defined by Waddington *et al.* in 1942 (*4*, *5*). Initially, it was defined to explain the mechanism of the gene expression change that was important to the cell differentiation at the developmental stage.

Today, DNA methylation, histone modification, and nucleosome are studied actively in the study of the epigenome. In particular, a lot of researches of DNA methylation are supported and conducted by research organizations, for example, the Cold Spring Harbor Laboratory, American Association for Cancer Research (AACR), the Gordon Research Conferences (GRC), the

Federation of American Societies for Experimental Biology (FASEB), and Keystone Symposia. In addition, large-scale cooperative researches are launched in Europe (The Networks of Excellence 'The Epigenome' and 'EpiGeneSys'), the USA (US National Institutes of Health (NIH) Roadmap Epigenomics Project and ENCODE), Canada, Asia, and worldwide (the International Human Epigenome Consortium (IHEC)). Therefore, the epigenome study, in particular, DNA methylation, is gathering attention from all over the world.

DNA methylation

In most of the DNA methylation studies, the researchers focus on the 5-methylcytosine in CpG dinucleotides. There are several types of research that report the methylation of non-CpG sequence in some species (*6*–*8*), however, which functions remain unknown.

Most of the effects of DNA methylation on gene expression are suppression (*9*). X-chromosome inactivation, imprinting, and some tissue-specific gene are common phenomena by DNA methylation alteration. Additionally, it is known that the DNA methylation changes in CpG-rich regions known as CpG islands (CGIs) in the transcriptionally regulated region, for example, promoter, enhancer, and insulator, are closely associated with gene expression.

The inducer of DNA methylation change

The DNA methylation change is induced by various factors. Tobacco smoke is one of the most popular inducers of DNA methylation change (*10*–*20*). In particular, it is revealed that smoking in pregnancy period is strongly associated with the global DNA methylation change in the fetus (*11*, *19*, *21*). Additionally, traffic-related air pollution is also correlated with DNA methylation alteration (*22*–*24*). Furthermore, aging is the key factor of global DNA methylation change (*25*, *26*).

DNA methylation change in cancer

In cancer, the abnormality of DNA methylation status has reported. It is revealed that mutations and/or deletions in tumor suppressor genes are related to cancer development. Additionally, in the recent cancer epigenome study, it is revealed that the alteration of DNA methylation in tumor suppressor gene reduce its expression, and induce cancer development (*27*). The study of epigenome alteration in various types of the tumor was performed, and its data are available in The Cancer Genome Atlas (https://cancergenome.nih.gov/).

Major bioresources for epidemiological studies

Because it is essential to use relatively large samples in searching for genes that are susceptible to multifactorial diseases, the DNA sources are limited to some cell types. The peripheral blood cells (PBCs) are one of the suitable cell types for analysis.

Epstein-Barr Virus (EBV) -transformed immortalized lymphoblastoid cell lines (LCLs) are also used to obtain DNA. EBV, also known as human herpesvirus 4 (HHV-4), were found by Epstein *et al.* (*28*) in the tumor cell from the patient of Burkitt's lymphoma. This finding is the first evidence of the tumor induced by viruses. EBV infection may also be associated with the development of the Hodgkin's disease, undifferentiated nasopharyngeal carcinoma, immunoblastic lymphomas arising in immunocompromised individuals, and T and NK cell lymphomas (*29*). LCLs can be generated from both healthy individuals and patients and supply an unlimited source of genomic DNA. Additionally, LCLs and PBCs have been successfully used for gene expression analyses (*30*).

However, it is known that DNA methylation status varies between cell types (*31*). Therefore, to extend our knowledge of the difference in DNA methylation status between LCLs and PBCs is important in human population studies that use these DNA sources to elucidate the epigenetic

risks for multifactorial diseases.

Study design

Figure 1 shows the general flow of this study. I designed experiments to compare the DNA methylation status between LCLs and PBCs at an epigenome-wide level using approximately 400,000 methylation data sites from 92 LCL and 192 PBC samples obtained using the Human Methylation 450K array. I analyzed global differences in methylation profiles and the degree of difference in methylation level of each site in terms of location (inside or outside the CpG island, the distance from transcription start site and promoter type) between LCLs and PBCs. Additionally, the association strength of methylation levels at the aging-related methylation sites in *FHL2* and *ELOVL2* with chronological age was compared between LCLs and PBCs.

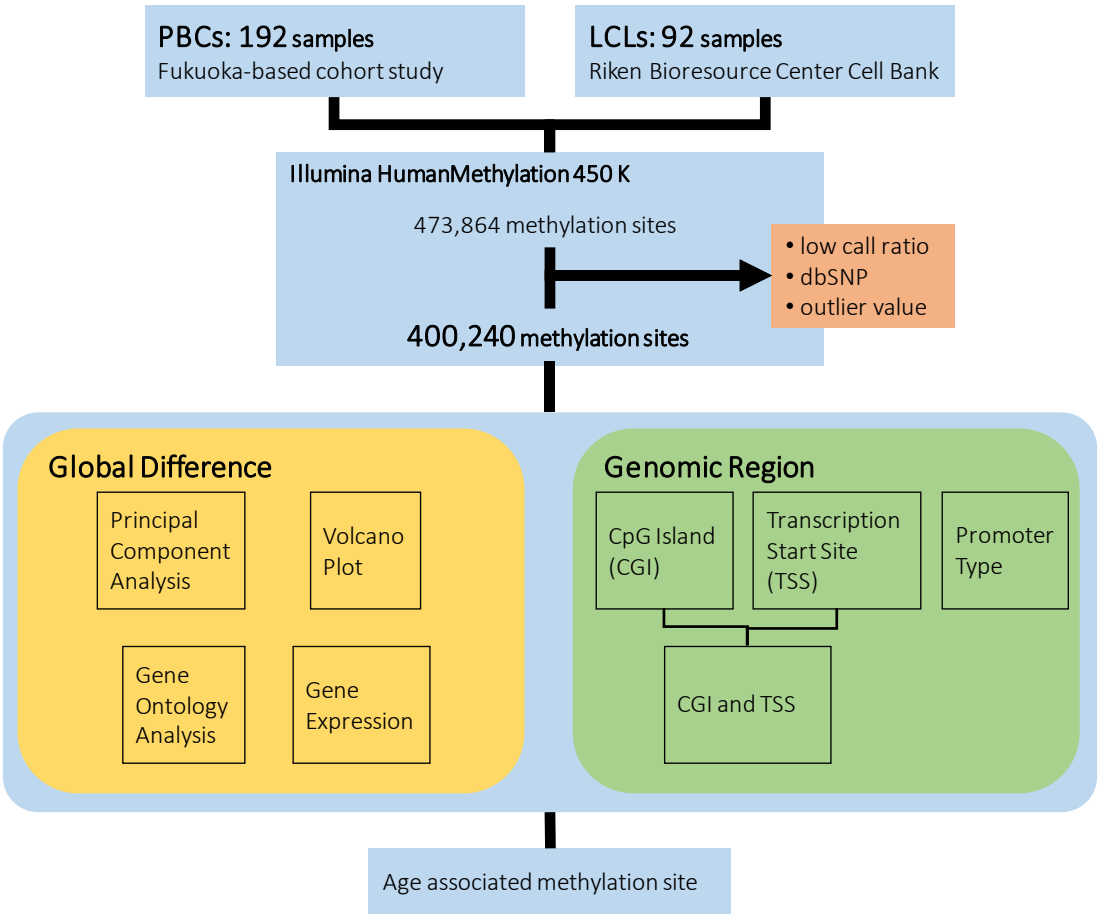


Figure 1: The study design.

Chapter 1: Methylation level measurement and methylation site selection

Subjects

EBV-transformed LCLs derived from 92 healthy Japanese subjects were provided by the Riken Bioresource Center Cell Bank (*32*). PBCs were obtained from 192 participants of a baseline survey of the general population from a Fukuoka-based cohort study (*33*, *34*). This study was performed in accordance with the principles of the Declaration of Helsinki and was approved by the Institutional Review Board at Kyushu University.

DNA methylation chip assay

Genomic DNA was bisulfite-treated using the EZ-96 DNA Methylation Kit (Zymo Research Corporation, Orange, CA), which combines bisulfite conversion and DNA cleanup in a 96-well plate. Genome-wide DNA methylation profiles were obtained using the Illumina HumanMethylation450 BeadChip (Illumina, San Diego, CA) according to the manufacturer's instructions. The GenomeStudio V2011.1 (Methylation Module version 1.9.0) was employed to determine the beta values that reflected the estimated methylation level for each CpG site. The beta value was calculated as: Max (signal for methylation, 0) / [Max (signal for methylation, 0) + Max (signal for unmethylation, 0) + 100]. Using this metric, the DNA methylation level was represented by a number between 0 (no methylation) and 1 (complete methylation). The signal intensities were normalized to the internal controls and background prior to beta value calculation.

Selection and classification of DNA methylation sites

The flowchart of methylation sites selection is shown in Figure 2. Among 473,864 methylation sites on the autosomes, 1,305 sites showing low calls (< 0.95) were removed for further analyses. To eliminate SNP-associated methylation sites, I screened the nearest SNP for each methylation site using the dbSNP135 database (SNPs categorized in weight = 1 group,

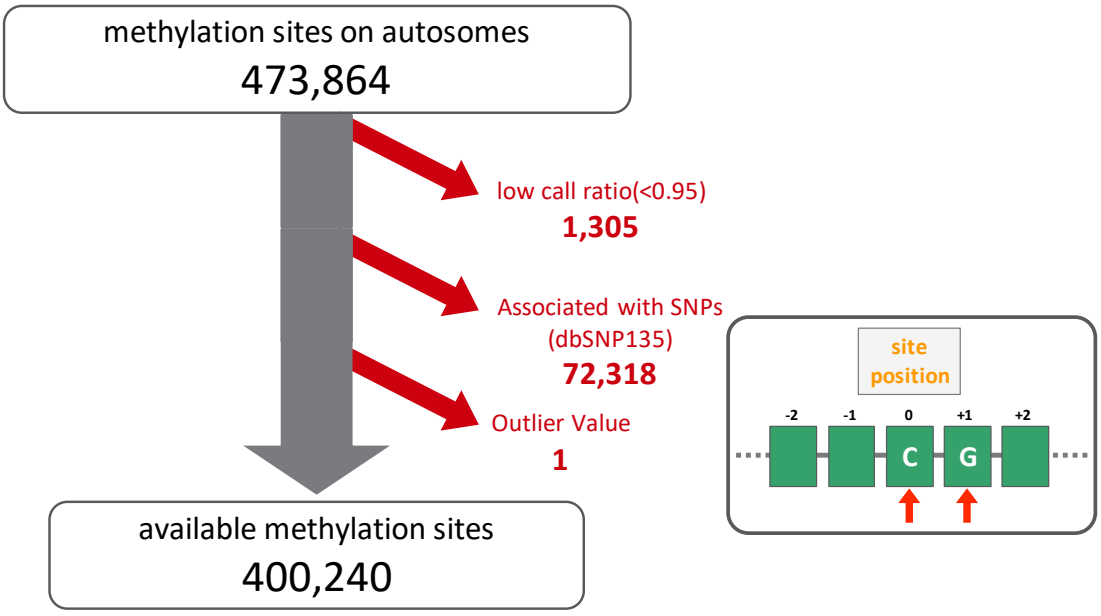


Figure 2: The selection of methylation sites.

The methylation sites showing low call ratio, associated with SNPs, and outlier value are removed in this study, and remained 400,240 sites are used in follow analysis.

http://www.ncbi.nlm.nih.gov/SNP/). I found 72,318 sites in which SNPs were located on the C or G site. Additionally, one methylation site demonstrated an outlier value. After removing these sites; 400,240 methylation sites on the array were available for further analyses.

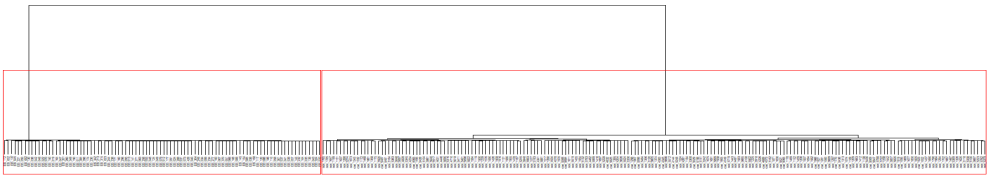
Statistical Analysis

To evaluate the difference in methylation level of each site, the data were analyzed using modeling individual Illumina beta values using a generalized linear model (glm) with cell type (LCLs or PBCs), age and sex as the independent variables. *P*-values and the difference in methylation level for each cell type were obtained. The statistical power to detect methylation differences of 0.25 and 0.5 between 192 PBCs and 92 LCLs was estimated to be 50.2% and 97.5%, respectively at a significance level of $P = 0.05$ using G*Power 3.1 software (*35*).

Chapter 2: Global difference between PBCs and LCLs

Cluster analysis

To assess the global difference of DNA methylation levels between LCLs and PBCs, I performed a hierarchical cluster analysis using the methylation data of 400,240 sites on autosomes obtained using the 450K methylation array. Figure 3 shows the results of hierarchical cluster analysis. The distance was calculated at Euclid distance and analyzed with the Ward method. Red rectangles show two major clusters. Sample names are shown as a list below the dendrogram. LCLs and PBCs were completely separated into different clusters by whole epigenome methylation status.



Cluster 1

1 EB_144	24 EB_271	47 EB_190	70 EB_023
2 EB_145	25 EB_384	48 EB_428	71 EB_011
3 EB_139	26 EB_396	49 EB_067	72 EB_012
4 EB_141	27 EB_234	50 EB_073	73 EB_024
5 EB_201	28 EB_238	51 EB_086	74 EB_031
6 EB_213	29 EB_342	52 EB_089	75 EB_036
7 EB_160	30 EB_349	53 EB_040	76 EB_037
8 EB_168	31 EB_277	54 EB_041	77 EB_032
9 EB_150	32 EB_296	55 EB_095	78 EB_035
10 EB_154	33 EB_137	56 EB_099	79 EB_048
11 EB_177	34 EB_164	57 EB_045	80 EB_049
12 EB_186	35 EB_100	58 EB_063	81 EB_047
13 EB_171	36 EB_104	59 EB_042	82 EB_055
14 EB_174	37 EB_117	60 EB_043	83 EB_057
15 EB_221	38 EB_120	61 EB_052	84 EB_066
16 EB_232	39 EB_079	62 EB_058	85 EB_230
17 EB_304	40 EB_082	63 EB_050	86 EB_288
18 EB_324	41 EB_124	64 EB_051	87 EB_070
19 EB_256	42 EB_130	65 EB_013	88 EB_315
20 EB_301	43 EB_110	66 EB_019	89 EB_370
21 EB_244	44 EB_112	67 EB_038	90 EB_034
22 EB_248	45 EB_085	68 EB_039	91 EB_046
23 EB_260	46 EB_106	69 EB_022	92 EB_091

Cluster 2

93 coe_1358	121 coe_1590	149 coe_1000	177 coe_1524	205 coe_5626	233 coe_5512	261 coe_5198
94 coe_1467	122 coe_1582	150 coe_1138	178 coe_1620	206 coe_5788	234 coe_5918	262 coe_5822
95 coe_1074	123 coe_1508	151 coe_1095	179 coe_1790	207 coe_5847	235 coe_5228	263 coe_5605
96 coe_1770	124 coe_1443	152 coe_1677	180 coe_1984	208 coe_5524	236 coe_5100	264 coe_5274
97 coe_1691	125 coe_1152	153 coe_1554	181 coe_1637	209 coe_5722	237 coe_5920	265 coe_5065
98 coe_1580	126 coe_1052	154 coe_1577	182 coe_1119	210 coe_5183	238 coe_5096	266 coe_5164
99 coe_1907	127 coe_1455	155 coe_1916	183 coe_1495	211 coe_5338	239 coe_5104	267 coe_5430
100 coe_1151	128 coe_1281	156 coe_1630	184 coe_1542	212 coe_5820	240 coe_5848	268 coe_5162
101 coe_1199	129 coe_1297	157 coe_1313	185 coe_1175	213 coe_4234	241 coe_5345	269 coe_5363
102 coe_1785	130 coe_1416	158 coe_1871	186 coe_1970	214 coe_4108	242 coe_5881	270 coe_5667
103 coe_1702	131 coe_1178	159 coe_1374	187 coe_1131	215 coe_4828	243 coe_5487	271 coe_5436
104 coe_1850	132 coe_1697	160 coe_1459	188 coe_1060	216 coe_4689	244 coe_5783	272 coe_5087
105 coe_1938	133 coe_1109	161 coe_1238	189 coe_4341	217 coe_4491	245 coe_5502	273 coe_5118
106 coe_1437	134 coe_1395	162 coe_1094	190 coe_4710	218 coe_4144	246 coe_5519	274 coe_5762
107 coe_1426	135 coe_1647	163 coe_1172	191 coe_4383	219 coe_4226	247 coe_5999	275 coe_5484
108 coe_1184	136 coe_1343	164 coe_1468	192 coe_4832	220 coe_4749	248 coe_5411	276 coe_5401
109 coe_1923	137 coe_1605	165 coe_1294	193 coe_4204	221 coe_4031	249 coe_5151	277 coe_5223
110 coe_1576	138 coe_1663	166 coe_1675	194 coe_4345	222 coe_5248	250 coe_5308	278 coe_5379
111 coe_1861	139 coe_1673	167 coe_1412	195 coe_4462	223 coe_4373	251 coe_5245	279 coe_5481
112 coe_1966	140 coe_1583	168 coe_1584	196 coe_4261	224 coe_5696	252 coe_5040	280 coe_5048
113 coe_1955	141 coe_1500	169 coe_1239	197 coe_4065	225 coe_5665	253 coe_5506	281 coe_5998
114 coe_1517	142 coe_1919	170 coe_1209	198 coe_4534	226 coe_5428	254 coe_5090	282 coe_5517
115 coe_1356	143 coe_1765	171 coe_1005	199 coe_4578	227 coe_5521	255 coe_5018	283 coe_5242
116 coe_1280	144 coe_1332	172 coe_1478	200 coe_4446	228 coe_5120	256 coe_5719	284 coe_5330
117 coe_1667	145 coe_1821	173 coe_1206	201 coe_5009	229 coe_5029	257 coe_5993	
118 coe_1225	146 coe_1937	174 coe_1611	202 coe_5689	230 coe_5132	258 coe_5684	
119 coe_1298	147 coe_1288	175 coe_1469	203 coe_5389	231 coe_5899	259 coe_5316	
120 coe_1326	148 coe_1869	176 coe_1167	204 coe_5857	232 coe_5825	260 coe_5861	

Figure 3: Hierarchical cluster analysis.

Upper part shows the dendrogram of the cluster analysis. Lower part shows the list of the samples classified in each cluster.

Principal component analysis

The results of the principal component analysis are shown in Figure 4A. The LCL and PBC groups were clearly distinguished by their first principal component score. In addition, the PBC samples were distributed within a narrow range, whereas the LCL samples showed a relatively wide range in the second principal component score. These results suggest that there is a global difference in DNA methylation levels between these cell types and that the levels are more diverse in LCLs than in PBCs.

Volcano plots

I then examined the difference in methylation level for each site using a glm adjusted for age and sex. As shown in the volcano plot in Figure 4B, the sites showing lower levels in LCL than in PBC were predominant (low-met-LCL group). The 138,871 sites (34.7% of the total) showed $-\log_{10}(P\text{-value}) > 10$; among these sites, 85.1% were in the low-met-LCL group. This inclination was observed in each autosome (Figure 5). Therefore, it was suggested that the main difference in DNA methylation between LCLs and PBCs was hypomethylation in the LCLs and that the change in methylation levels occurred globally in the autosomes.

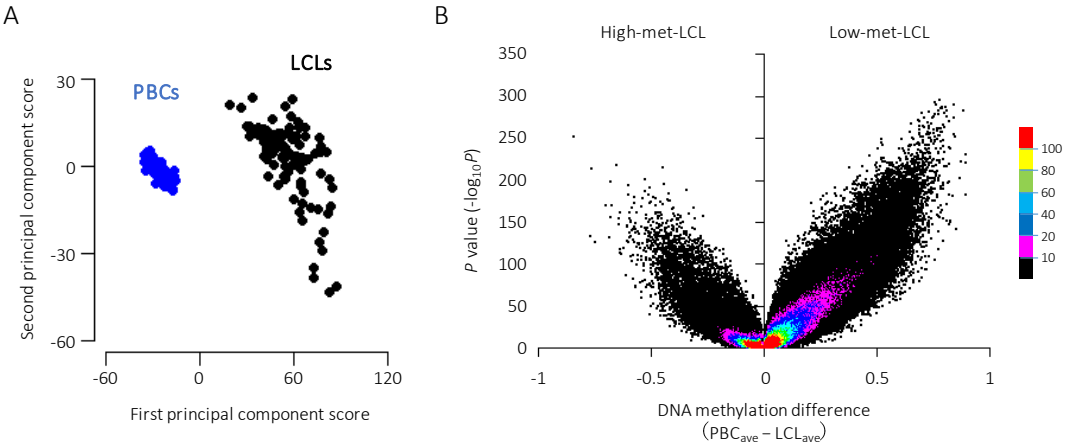


Figure 4: Global difference in the DNA methylation level between the LCLs and PBCs

(A) Principal component analysis (PCA) plot. PCA was performed using the methylation level of the 400,240 sites on autosomes. The LCL and PBC samples are shown in black and blue dots, respectively. (B) Volcano plot with the difference of the average of DNA methylation level on the x-axis and the *P*-value ($-\log_{10}$(*P*-value)) obtained via glm analysis on the y-axis. Each color shows the dot density ($100 < n$, $80 < n \leq 100$, $60 < n \leq 80$, $40 < n \leq 60$, $20 < n \leq 40$, $10 < n \leq 20$ and $n \leq 10$ per unit area (0.002 × 1 for x and y-axis, respectively) in red, yellow, green, sky blue, blue, pink and black, respectively).

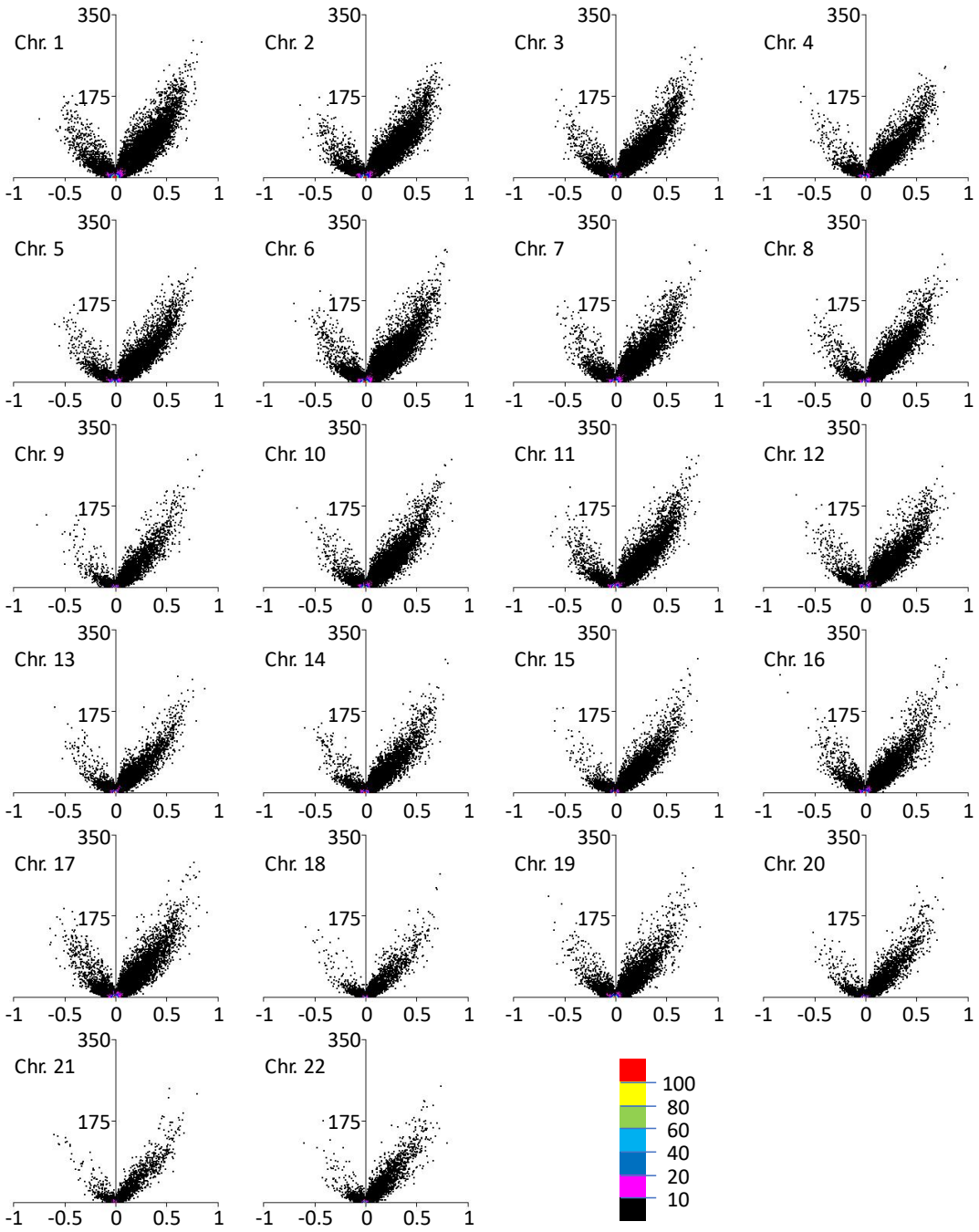


Figure 5: Volcano plot for each autosomal chromosome.

Volcano plot for each autosome with the difference of the average of DNA methylation levels on the x-axis and the *P*-value ($-\log_{10}$(*P*-value)) obtained via glm analysis on the y-axis. Each color shows the dot density ($100 < n$, $80 < n \leq 100$, $60 < n \leq 80$, $40 < n \leq 60$, $20 < n \leq 40$, $10 < n \leq 20$ and $n \leq 10$ per unit area (0.002×1 for the x-axis and y-axis, respectively) in red, yellow, green, sky blue, blue, pink and black, respectively).

Gene ontology analysis

Additionally, to investigate the association between DNA methylation change and gene biological function, I performed gene ontology enrichment analysis. In this study, I annotated reference gene name from UCSC Genome Bioinformatics database (http://genome.ucsc.edu/index.html) to the 6,689 methylation sites located inside of gene region and have a significant difference of methylation level between PBCs and LCLs ($-\log_{10}$(*P*-value) > 100). The 3,779 genes were annotated to the methylation sites. The details of the result are shown in Table 1.

Then, I performed gene ontology (GO) enrichment analysis with DAVID (https://david.ncifcrf.gov/home.jsp), and the results in all sites, high-met-LCL, and low-met-LCL are shown in Supplementary Table 1, Supplementary Table 2, and Table 2, respectively. It is revealed that the GO terms associated with development are significantly enriched in low-met-LCL group (GO:0007275 (multicellular organism development), *P*-value of 1.27×10^{-34}; GO:0048731 (system development), *P*-value of 5.89×10^{-32}; GO:0048856 (anatomical structure development), *P*-value of 4.43×10^{-29}; GO:0044767 (single-organism developmental process), *P*-value of 5.27×10^{-29}; GO:0032502 (developmental process), *P*-value of 9.93×10^{-29}).

Table 1: Gene ontology analysis

			GO Term						
	site	gene	DAVID ID	Biological Process	%	Cellular Component	%	Molecular Function	%
All	6,689	3,779	3,524	2,973	84.4	3,146	89.3	2,936	83.3
High-met-LCL	568	427	408	366	89.7	385	94.4	378	92.6
Low-met-LCL	6,121	3,448	3,208	2,694	84.0	2,851	88.9	2,644	82.4

Table 2: GO term enrichment analysis in low-met-LCL

Term (Biological Process)	Description	Count	%	*P*-Value	FDR
GO:0044707	single-multicellular organism process	1,241	38.7	4.33E-36	8.80E-33
GO:0007275	multicellular organism development	1,048	32.7	9.49E-34	1.93E-30
GO:0032501	multicellular organismal process	1,409	43.9	2.75E-33	5.58E-30
GO:0048731	system development	938	29.2	4.24E-32	8.61E-29
GO:0048856	anatomical structure development	1,134	35.3	4.81E-30	9.77E-27
GO:0044767	single-organism developmental process	1,133	35.3	9.27E-30	1.88E-26
GO:0032502	developmental process	1,158	36.1	2.12E-29	4.30E-26
GO:0030154	cell differentiation	797	24.8	5.74E-26	1.17E-22
GO:0023052	signaling	1,236	38.5	5.50E-24	1.12E-20
GO:0044699	single-organism process	2,329	72.6	9.09E-24	1.85E-20
Term (Cellular Component)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0071944	cell periphery	1,168	36.4	3.76E-52	6.01E-49
GO:0005886	plasma membrane	1,141	35.6	3.71E-50	5.93E-47
GO:0044459	plasma membrane part	653	20.4	3.14E-40	5.01E-37
GO:0045202	synapse	243	7.57	6.75E-31	1.08E-27
GO:0097458	neuron part	367	11.4	7.46E-31	1.19E-27
GO:0005887	integral component of plasma membrane	414	12.9	1.43E-26	2.29E-23
GO:0031226	intrinsic component of plasma membrane	426	13.3	2.58E-26	4.12E-23
GO:0044456	synapse part	192	5.99	1.69E-23	2.70E-20
GO:0043005	neuron projection	262	8.17	5.05E-21	8.07E-18
GO:0030054	cell junction	344	10.7	7.10E-21	1.13E-17
Term (Molecular Function)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0022836	gated channel activity	122	3.8	9.34E-21	1.63E-17
GO:0022838	substrate-specific channel activity	148	4.61	1.60E-20	2.79E-17
GO:0005216	ion channel activity	143	4.46	5.88E-20	1.03E-16
GO:0046873	metal ion transmembrane transporter activity	141	4.4	8.58E-20	1.50E-16
GO:0005261	cation channel activity	112	3.49	3.32E-19	5.79E-16
GO:0022803	passive transmembrane transporter activity	152	4.74	6.51E-19	1.14E-15
GO:0015267	channel activity	151	4.71	1.35E-18	2.36E-15
GO:0015075	ion transmembrane transporter activity	209	6.51	8.37E-14	1.46E-10
GO:0005244	voltage-gated ion channel activity	72	2.24	1.56E-13	2.72E-10
GO:0022832	voltage-gated channel activity	72	2.24	1.56E-13	2.72E-10

Association with gene expression and DNA methylation

I also investigated the association with DNA methylation difference and gene expression. In this analysis, I used the gene expression data reported by Powell *et al.*(*36*). The association with DNA methylation difference between PBCs and LCLs and the changes of gene expression in the differentially methylated sites ($-\log_{10}$(*P*-value) > 100) are shown in Figure 6. It is revealed that there is no positive correlation between methylation level and gene expression.

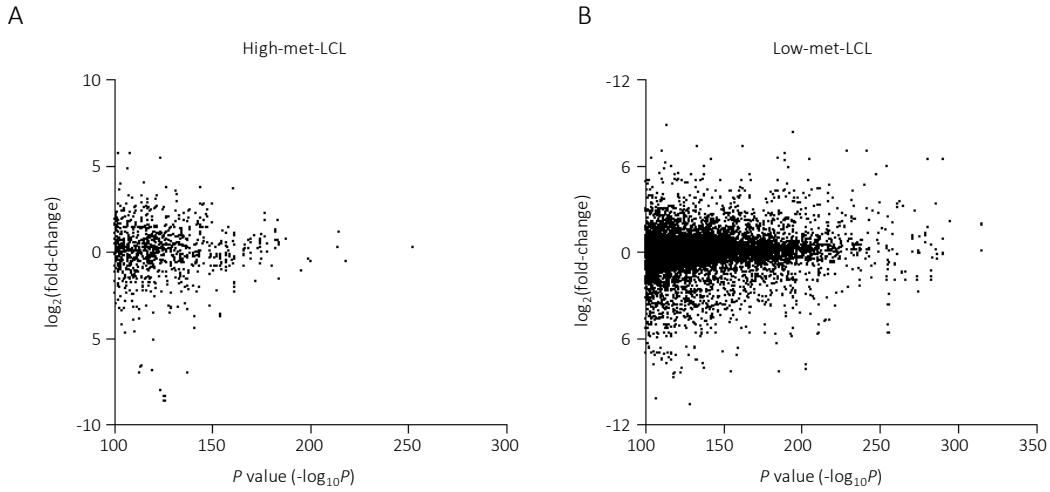


Figure 6: The association with DNA methylation difference and gene expression in genetic region.

Scatter plot with the *P*-value ($-\log_{10}$(*P*-value)) obtained via glm analysis of DNA methylation difference between PBCs and LCLs on x-axis, and the ratio of gene expression change according to EB virus infection ($\log_2$(gene expression (EBV^+) / gene expression (uninfected))) are on y-axis. The results in high-met-LCL group and low-met-LCL group are shown in Figure 6A and Figure 6B, respectively.

It is already known that the DNA methylation in promoter region has an effect on the gene expression (*9*). To investigate the correlation between methylation level and gene expression in detail, I performed a similar analysis in the methylation sites located in the promoter region (-500 bp < TSS < 2000 bp). In Figure 7, there is a slight correlation between DNA methylation changes and gene expression.

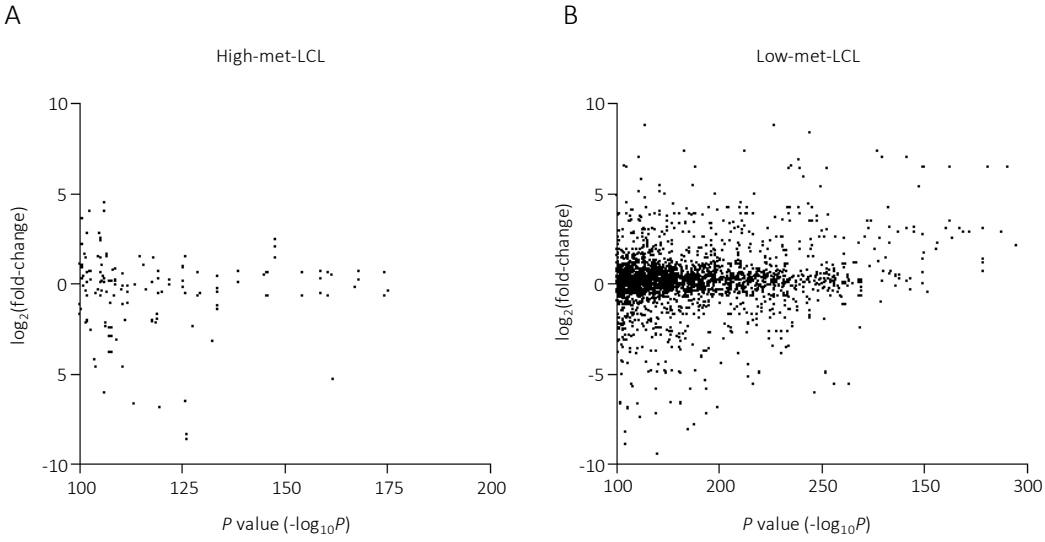


Figure 7: The association with DNA methylation difference and gene expression in promoter region.

Scatter plot with the *P*-value (-$\log_{10}$(*P*-value)) obtained via glm analysis of DNA methylation difference between PBCs and LCLs on x-axis, and the ratio of gene expression change according to EB virus infection ($\log_2$(gene expression (EBV^+) / gene expression (uninfected))) are on y-axis. The results in high-met-LCL group and low-met-LCL group are shown in Figure 7A and Figure 7B, respectively.

Chapter 3: Association with CpG islands

The annotation of CpG islands

Based on the CpG Islands (CGI) track of the UCSC table browser of the UCSC Genome Bioinformatics database (http://genome.ucsc.edu/index.html), the 400,240 sites on autosomes were classified into two groups, CGI-sites (135,674 sites, inside of CGI) or non-CGI-sites (264,566 sites, outside the CGI). Among the non-CGI sites, 95,625 sites were located near CGI (±2,000 bases) that were classified in a shore group.

CpG island and non-CpG island

I next assessed the distribution of the difference in methylation levels between LCLs and PBLs in terms of the location of the site (inside or outside the CpG island) (named CGI-site or non-CGI-site). As shown in Figure 8A, the distribution of difference was dissimilar between them; the proportion of the sites showing a low *P*-value was larger in the non-CGI-site group (black solid line) than in the CGI-site group (black dashed line). This trend was apparent in the low-met-LCL group (compare the red solid and dashed lines), whereas a dissimilarity of distribution was not observed in the high-met-LCL group (compare the blue solid and dashed lines). These results prompted us to further classify the non-CGI-sites into shore or non-shore groups because the CGI shores were suggested to contribute to tissue-specific DNA methylation (*37*, *38*). However, I did not find significant differences in the distribution between the shore and non-shore group of the low-met-LCL (Figure 8B). Taken together, these results suggested that the majority of hypomethylation observed in the LCLs occurred at sites outside the CGIs regardless of shores.

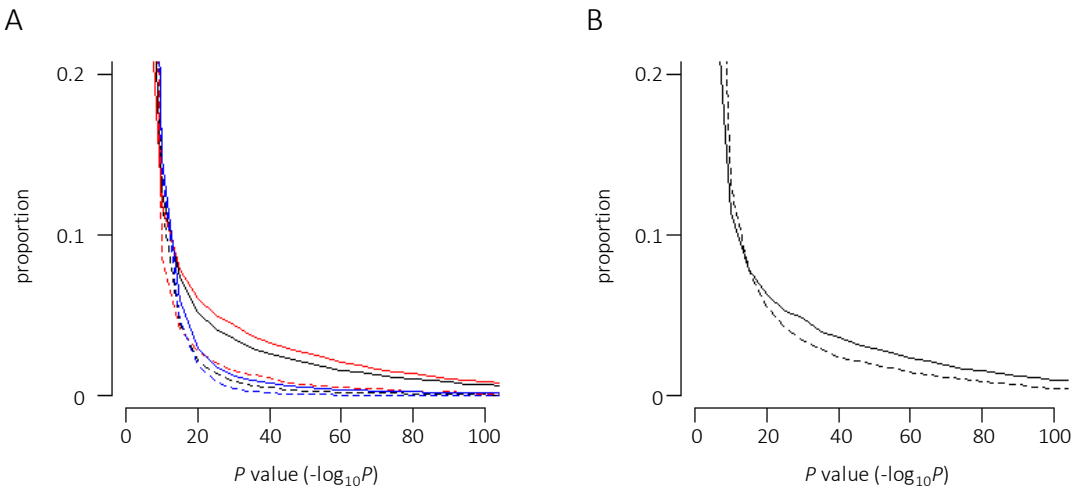


Figure 8: Distribution of the differences in methylation levels between LCLs and PBLs in terms of CGI.

(A) The proportion of P-values obtained from non-CGI and CGI sites in all samples (black solid and dashed lines, respectively), non-CGI and CGI sites in the low-met-LCL group (red solid and dashed lines, respectively), and non-CGI and CGI sites in the high-met-LCL group (blue solid and dashed lines, respectively) are indicated. (B) The proportion of P-values obtained from the non-shore and shore sites (solid and dashed lines, respectively) in the non-CGI sites of the low-met-LCL group are indicated.

Chapter 4: Association with distance from transcription start site

Transcription start site position

The distance between the methylation site and the nearest transcription start site (TSS) was calculated using the NCBI RefSeq database. The physical positions on the human genome were based on the Genome Reference Consortium Human Build 37 (GRCh37, http://www.ncbi.nlm.nih.gov/assembly/).

Distance from TSS and DNA methylation difference

I further examined the relationship between the distance from the TSS and the difference in DNA methylation levels observed among LCLs and PBCs. I plotted $-\log_{10}$(*P*-value) for each site against the distance from the nearest TSS (shown in gray dots in Figure 9) and indicated a proportion of the site showing $-\log_{10}$(*P*-value) > 10, 25 and 50 in blue, green and pink dots, respectively (Figure 9). The proportion was calculated by dividing the number of the sites meeting the *P*-value criteria by the total number of sites within ±50 bases of window size. I found that the proportion of significantly different sites was lower near the TSS. For instance, approximately 25% of the sites near the TSS showed $-\log_{10}$(*P*-value) > 10, whereas this proportion increased to approximately 45% for the sites located approximately ±1,000 bases from the TSS in the low-met-LCL group (blue dots, Figure 9A). This trend was also observed even in the lower *P*-value threshold group (green and pink dots) and in the high-met-LCL group (Figure 9B).

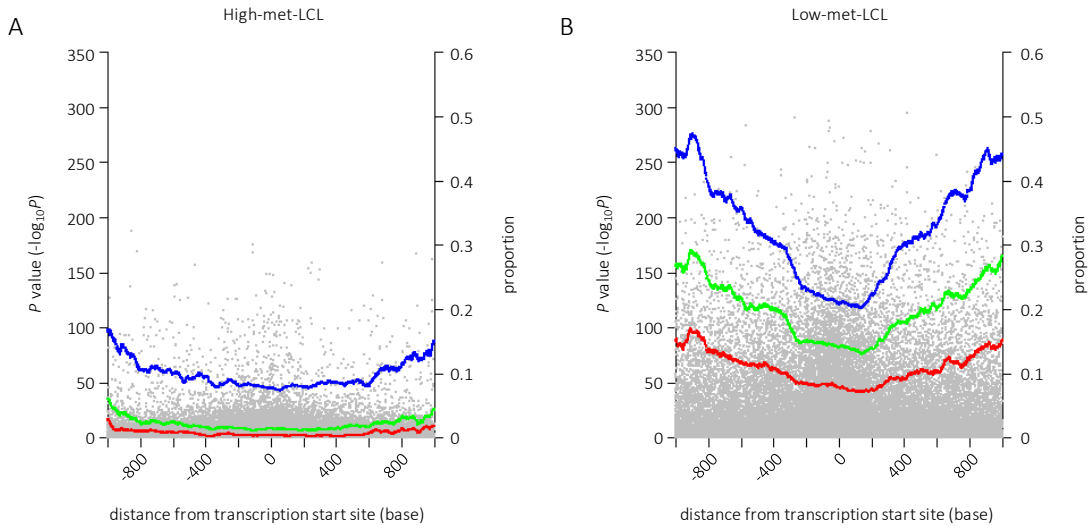


Figure 9: Distribution of the differences in methylation levels between LCLs and PBLs in terms of TSS.

P-values were plotted against the distance from the nearest TSS (gray dots). The proportion of the sites with *P*-values ($-\log_{10}$(*P*-value)) greater than 10 (blue dots), 25 (green dots) and 50 (pink dots) in a window size of ±50 bases were plotted. Figure 9A and Figure 9B show the results in the high-met-LCL group and the low-met-LCL group, respectively.

Synergistic action of CGI and TSS

I then analyzed the sites showing $-\log_{10}$(*P*-value) > 10 separately for CGI- and non-CGI-site groups. As shown in Figure 10, the proportion of non-CGI-sites near the TSS was high in both the low- and high-met-LCL groups (red and blue dots, respectively, Figure 10). However, the lowest proportion was observed near the TSS in the case of CGI-sites (orange and sky-blue dots for low- and high-met-LCL groups, respectively, Figure 10). These results suggested that the low CpG promoter would show a more significant difference in DNA methylation levels than the high CpG promoter.

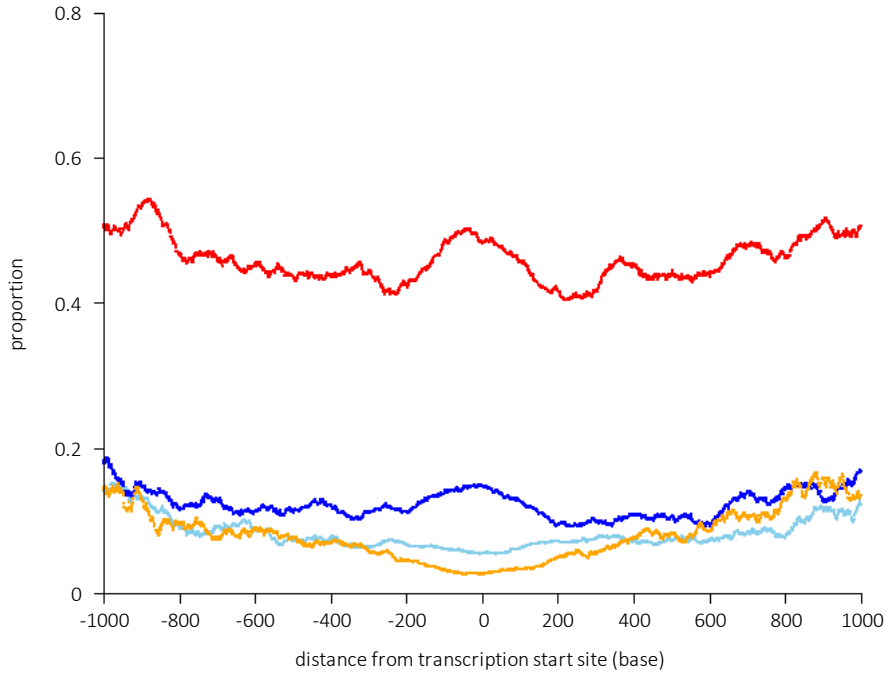


Figure 10: Synergistic effect of CGI and TSS

The proportion of the sites with *P*-values ($-\log_{10}$(*P*-value)) greater than 10 obtained from non-CGI and CGI sites in the low-met-LCL group (red and orange dots, respectively), and from non-CGI and CGI sites in the high-met-LCL group (blue and sky-blue dots, respectively) in a window size of ±50 bases were plotted against TSS.

Chapter 5: Association with promoter type

Definition of promoter type

Of 400,240 probes, 159,688 demonstrated a TSS between -500 bases and +2,000 bases; among these, 85,700 sites could be classified into high-CpG-density promoters (HCP), intermediate-CpG-density promoters (ICP) and low-CpG-density promoters (LCP), as reported by Mikkelsen *et al*. (*39*) (69,836, 10,719 and 5,145 in HCP, ICP and LCP, respectively).

Differentially methylated sites in promoter region

I analyzed the distribution of $-\log_{10}$(*P*-value) in all, low- and high-met-LCL groups and results are shown in Figure 11. It was shown that the proportion of differentially methylated sites was higher in the LCPs than the HCPs. In the LCPs, the proportion of the sites showing $-\log_{10}$(*P*-value) > 25 was 30.7%, whereas that in HCPs was 4.1% in all sites (compare Figure 11A and Figure 11G). This was more pronounced in the low-met-LCL group (compare Figure 11B, Figure 11C, Figure 11H and Figure 11I). The sites located in ICPs showed intermediate values between HCPs and LCPs (Figure 11D, Figure 11E, and Figure 11F). These results suggested that the methylation sites located in low CpG promoters could be sensitive to demethylation in LCLs.

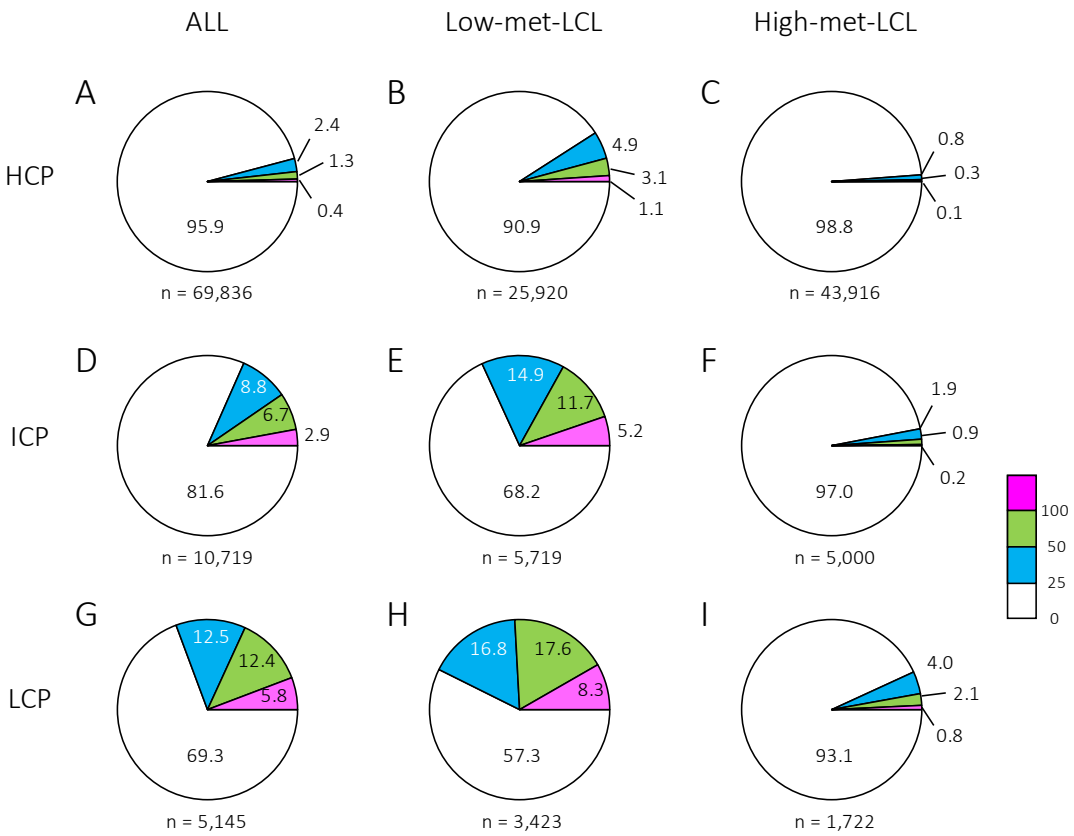


Figure 11: Difference in methylation levels between LCLs and PBLs in terms of promoter type.

The proportion of the sites with *P*-values ($-\log_{10}$(*P*-value)) ≤ 25, 25-50, 50-100 and ≥ 100 are indicated in white, blue, green and pink, respectively. The results obtained from the HCP, ICP and LCP sites in all samples (A, D and G, respectively) in the low-met-LCL group (B, E and H, respectively) and in the high-met-LCL (C, F and I, respectively) are shown.

To further assess promoter type differences, I compared the HCPs, ICPs, and LCPs methylation level profiles. As shown in Figure 12, nearly half of the sites in ICPs and LCPs showed more than 0.6 methylation levels, whereas almost all sites in HCPs were hypomethylated in PBCs. Additionally, it was observed that the methylation levels of highly methylated sites of the LCPs decreased in the LCLs. Therefore, I concluded that highly methylated sites of LCPs caused the difference in DNA methylation levels observed between HCPs and LCPs, especially in the low-met-LCL group.

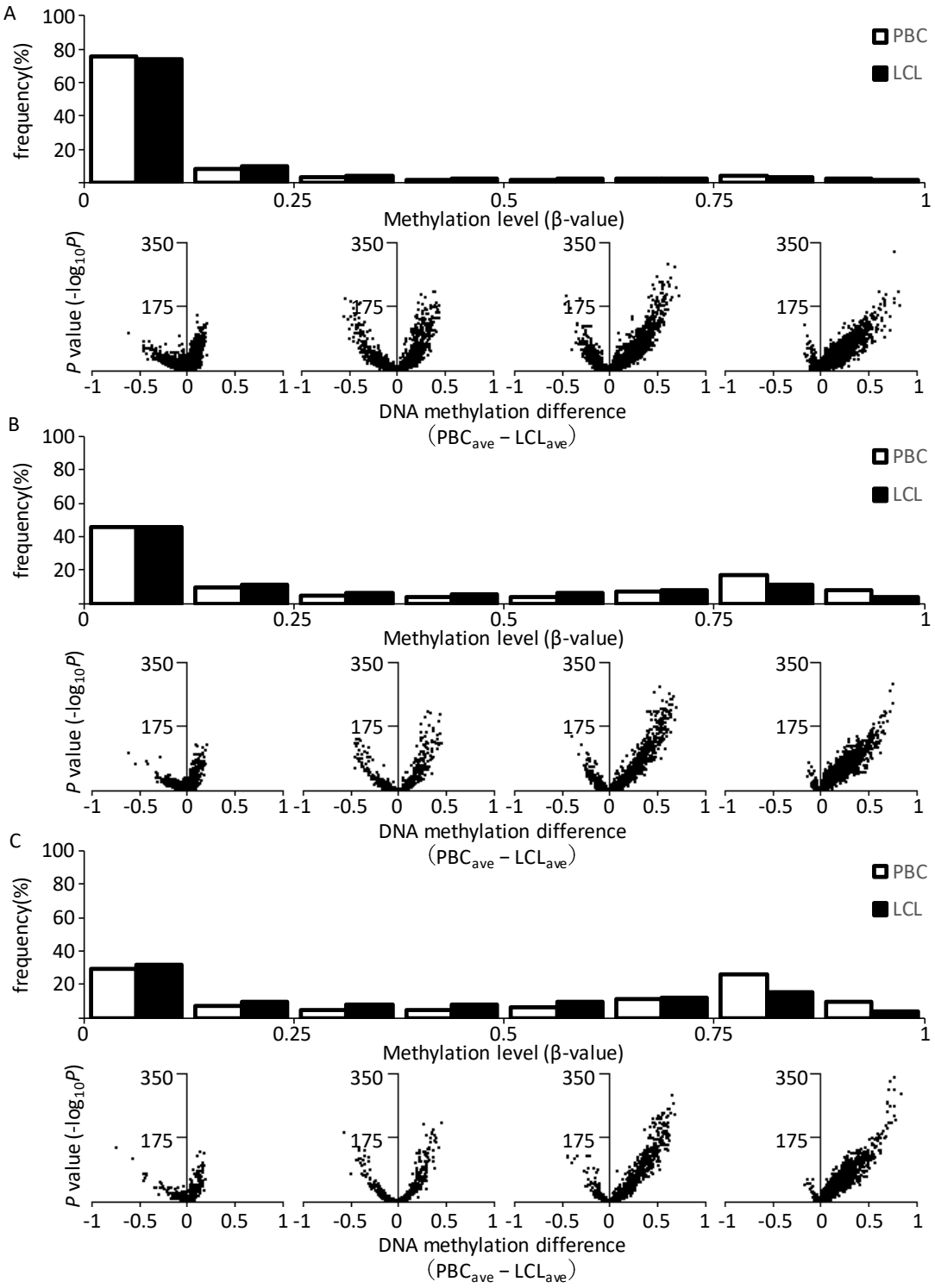


Figure 12: Distribution of the methylation levels of the sites in HCPs, ICPs, and LCPs.

The results in HCPs, ICP, and LCPs are shown in A, B, and C, respectively.

(Upper panel) The distribution of the methylation levels of the sites.

(legend continued on next page)

(Lower panel) Volcano plot with the difference of the average of DNA methylation level on the x-axis and the *P*-value ($-\log_{10}$(*P*-value)) obtained via glm analysis on the y-axis. Each plot shows the sites with methylation level of 0-0.25, 0.25-0.5, 0.5-0.75 and 0.75-1 in PBCs.

Chapter 6: The methylation level difference in age-associated methylation site

Age-associated methylation sites

Using DNA obtained from PBCs, it has been reported that the methylation levels of several CpG sites are associated with chronological age. However, it remains unclear whether LCLs should be utilized for studies on epigenetic aging biomarkers. To address this issue, I performed a regression analysis for chronological age and known aging-related CpG sites located in *FHL2* and *ELOVL2* (*25*, *26*). *FHL2* encodes a member of the four-and-a-half-LIM-only protein family that is suggested to have a role in the assembly of extracellular membranes and in the transformation of normal myoblasts to rhabdomyosarcoma cells (OMIM 602633). *ELOVL2* encodes an enzyme that catalyzes the first and rate-limiting reaction of the long-chain fatty acids elongation cycle (OMIM 611814). As shown in Figure 13, the methylation level of the PBCs was highly correlated with chronological age (blue dots, P = 1.7E-18 and r^2 = 0.33 for *FHL2*, P = 3.1E-25 and r^2 = 0.44 for *ELOVL2*). In contrast, the methylation level of the LCLs was varied and the association was weak (black dots, P = 0.04 and r^2 = 0.05 for *FHL2*, P = 1.9E-5 and r^2 = 0.18 for *ELOVL2*). Therefore, these results suggest that DNA obtained from LCLs may not always be an alternative to DNA from PBCs.

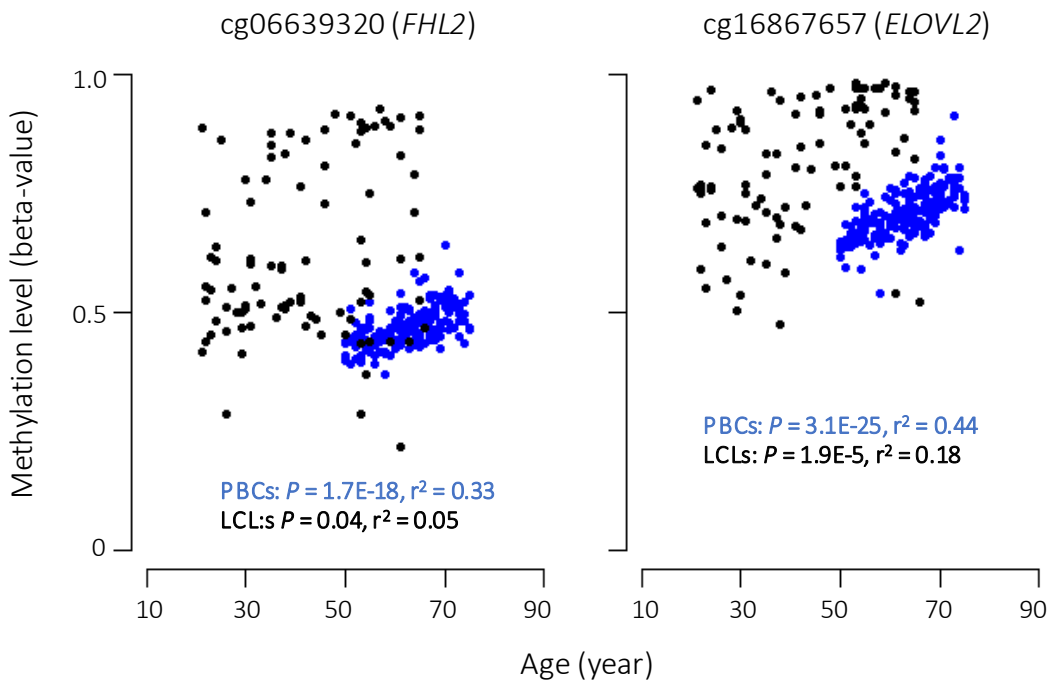


Figure 13: Regression analyses of the methylation levels and chronological age at the *FHL2* and *ELOVL2* loci.

The methylation levels in the LCLs (black dots) and PBCs (blue dots) were plotted against the age of the donors at the time of providing the specimens. The P-values and r2 were obtained by correcting for sex.

Discussion

In this study, I used a 450K methylation array to investigate the methylation differences between LCLs and PBCs, which are commonly used in genetic epidemiological studies. In all genomes, the majority of the sites in the LCLs showed lower methylation levels than those of the PBCs, and these sites were primarily located in non-CGI regions. Additionally, I found that differentially methylated sites were predominantly located in the LCP region.

In the differentially methylated sites, I found that GO terms that associated with development are enriched. Such tendency was not seen in the list of genes known as oncogene reported in Futreal *et al.* (*40*). On the other hand, there are such trend in the top 500 genes differentially expressed in EBV-infected cells (*41*), therefore it is suggested that this tendency may be caused by EBV infection.

I investigated the correlation of DNA methylation and gene expression, and I found that there was no significant correlation in the methylation sites located in the gene region. However, if I limited to the methylation sites located in the promoter region, there is a slight correlation. These results suggest that DNA methylation changes are not directly associated with gene expression change, however, there may be some interactions with other various factors.

Although a relatively small sample number and a number of methylation sites were analyzed, previous studies showed that methylation status in LCLs is different from that of PBCs and that the methylation level in LCLs is lower than that of PBCs in the majority of sites (*42–47*). Because a large number of samples and more sites were examined, I could investigate the differences in methylation levels between LCLs and PBCs in terms of CGI location, distance from TSS and promoter type as characterized by CG density. I found that a fraction showing a significant difference in methylation level between the LCLs and PBCs was observed near the TSS in the non-CGI sites but not in the CGI sites. This result suggests that the difference in the methylation

level of these cell types would be high in the genes in which the promoter shows a low GC content.

I found that significantly different methylation sites were predominant in LCPs but not in HCPs. It has been demonstrated that LCPs are generally associated with tissue-specific genes, whereas HCPs are associated with two classes of genes, including ubiquitous 'housekeeping' genes and highly regulated 'key developmental' genes (*39*, *48*, *49*). Therefore, my results suggest that the methylation sites located in promoters classified as LCP could have a functional role in distinguishing between LCLs and PBCs by regulating the corresponding gene expression.

The epigenome-wide association studies using human population samples to identify the disease risk loci and epigenomes that are affected by intrinsic or extrinsic factors, such as aging and smoking, have been progressing (*14*, *15*, *25*, *26*). I evaluated the differences in association strength between well-known aging methylation sites and the chronological age of the samples between LCLs and PBCs and found that the correlation was more significant in PBCs than LCLs. This was due to a larger variance of methylation levels in LCLs than in PBCs. In addition to the differences in cell type, artificial experimental processes, including in vitro culture, culture period and culture freezing and thawing could cause the large variances in data observed in the LCLs. Therefore, I concluded that DNA obtained from LCLs may not always be a proxy for DNA from PBCs in studies of epigenome-wide analysis attempting to elucidate the role of epigenetic change in disease risks.

Conclusion

There is a global difference in DNA methylation levels between LCLs and PBCs, and the main difference was hypomethylation in the LCLs. The methylation levels of highly methylated sites of the low-CpG-density promoters in PBCs decreased in the LCLs, suggesting that the methylation sites located in low-CpG-density promoters could be sensitive to demethylation in LCLs. The correlation between well-known ageing methylation sites and the chronological age

of the samples was more significant in PBCs than LCLs, indicating that despite being generated from a single cell type, LCLs may not always be a proxy for DNA from PBCs in studies of epigenome-wide analysis attempting to elucidate the role of epigenetic change in disease risks.

Acknowledgements

I would like to express my gratitude to Prof. Ken Yamamoto, Department of Medical Biochemistry, Kurume University School of Medicine, for giving many important and helpful advice for this thesis. I am deeply grateful to Assoc. Prof. Hiroki Shibata, Division of Genomics, Medical Institute of Bioregulation, Kyushu University, for supervising this thesis. I would like to thank Dr. Keizo Ohnaka, Department of Geriatric Medicine, Graduate School of Medical Sciences, Kyushu University, for collecting samples. I also thank Ms. Chihiro Iwaya, Division of Genomics, Medical Institute of Bioregulation, Kyushu University, for statistical and bioinformatics analysis. I thank all of the people who have continuously supported the population-based cohort study, the Kyushu University Fukuoka Cohort Study. I also thank Ms. Miki Sonoda for her technical assistance. Finally, I want to thank my parents and my wife Reina for warm encouragements and supports.

Funding

This work was supported by KAKENHI Grant Number 15K08290 from the Japan Society for the Promotion of Science.

Abbreviations

LCL: lymphoblastoid cell line; PBC: peripheral blood cell; CGI: CpG Island; TSS: Transcription start site; HCP: high-CpG-density promoter; ICP: intermediate-CpG-density promoter; LCP: low-CpG-density promoter; glm: generalized linear model; PCA: principal component analysis; *FHL2*: Four and a half LIM domains 2; *ELOVL2*: Elongation of very long chain fatty acids protein 2

References

1. J. M. Ordovás, C. E. Smith, Epigenetics and cardiovascular disease. *Nat. Rev. Cardiol.* **7**, 510–9 (2010).
2. K. H. Costenbader, S. Gay, M. E. Alarcón-Riquelme, L. Iaccarino, A. Doria, Genes, epigenetic regulation and environmental factors: which is the most relevant in developing autoimmune diseases? *Autoimmun. Rev.* **11**, 604–9 (2012).
3. S. T. Keating, A. El-Osta, Epigenetic changes in diabetes. *Clin. Genet.* **84**, 1–10 (2013).
4. C. H. Waddington, The epigenotype. 1942. *Int. J. Epidemiol.* **41**, 10–13 (2012).
5. C. H. WADDINGTON, Canalization of Development and the Inheritance of Acquired Characters. *Nature*. **150**, 563–565 (1942).
6. R. Lister *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. **462**, 315–322 (2009).
7. S. J. Cokus *et al.*, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*. **452**, 215–219 (2008).
8. M. R. Rountree, E. U. Selker, DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. *Genes Dev.* **11**, 2383–2395 (1997).
9. P. A. Jones, Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
10. Y. Zhang *et al.*, F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int. J. Epidemiol.* **43**, 1215–1225 (2014).
11. C. V. Breton *et al.*, Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS One*. **9**, e99716 (2014).
12. A. J. Drake *et al.*, In utero exposure to cigarette chemicals induces sex-specific disruption of one-carbon metabolism and DNA methylation in the human fetal liver.

BMC Med. **13**, 18 (2015).

13. R. V Steenaard *et al.*, Tobacco smoking is associated with methylation of genes related to coronary artery disease. *Clin. Epigenetics*. **7**, 54 (2015).

14. E. S. Wan *et al.*, Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.* **21**, 3073–82 (2012).

15. L. P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.* **88**, 450–7 (2011).

16. Q. Tan *et al.*, Epigenomic analysis of lung adenocarcinoma reveals novel DNA methylation patterns associated with smoking. *Onco. Targets. Ther.* **6**, 1471–9 (2013).

17. X. Gao, Y. Zhang, L. P. Breitling, H. Brenner, Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. *Oncotarget*. **7**, 46878–46889 (2016).

18. L. K. Küpers *et al.*, DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int. J. Epidemiol.* **44**, 1224–1237 (2015).

19. J. Z. J. Maccani, D. C. Koestler, E. A. Houseman, C. J. Marsit, K. T. Kelsey, Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. *Epigenomics*. **5**, 619–30 (2013).

20. S. A. Belinsky *et al.*, Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers. *Cancer Res.* **62**, 2370–2377 (2002).

21. F. a Taki, X. Pan, M.-H. Lee, B. Zhang, Nicotine exposure and transgenerational impact: a prospective study on small regulatory microRNAs. *Sci. Rep.* **4**, 7513 (2014).

22. C. Yauk *et al.*, Germ-line mutations, DNA damage, and global hypermethylation in mice

exposed to particulate air pollution in an urban/industrial location. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 605–10 (2008).

23. R. Ding *et al.*, Characteristics of DNA methylation changes induced by traffic-related air pollution. *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.* **796**, 46–53 (2016).

24. A. Baccarelli *et al.*, Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.* **179**, 572–578 (2009).

25. P. Garagnani *et al.*, Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*. **11**, 1132–1134 (2012).

26. G. Hannum *et al.*, Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*. **49**, 359–367 (2013).

27. T. Ushijima, Detection and interpretation of altered methylation patterns in cancer cells. *Nat. Rev. Cancer*. **5**, 223–231 (2005).

28. M. A. EPSTEIN, B. G. ACHONG, Y. M. BARR, Virus Particles in Cultured Lymphoblasts From Burkitt'S Lymphoma. *Lancet (London, England).* **1**, 702–3 (1964).

29. L. S. Young, A. B. Rickinson, Epstein–Barr virus: 40 years on. *Nat. Rev. Cancer*. **4**, 757–768 (2004).

30. J. E. Powell *et al.*, Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22**, 456–66 (2012).

31. M. J. Ziller *et al.*, Charting a dynamic DNA methylation landscape of the human genome. *Nature*. **500**, 477–81 (2013).

32. M. Iwakawa *et al.*, DNA repair capacity measured by high throughput alkaline comet assays in EBV-transformed cell lines and peripheral blood cells from cancer patients and healthy volunteers. *Mutat. Res.* **588**, 1–6 (2005).

33. A. Nanri *et al.*, Dietary patterns and C-reactive protein in Japanese men and women. *Am.*

J. Clin. Nutr. **87**, 1488–96 (2008).

34. D. Yoshida *et al.*, Waist circumference and cardiovascular risk factors in Japanese men and women. *J. Atheroscler. Thromb.* **16**, 431–41 (2009).

35. F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*. **39**, 175–91 (2007).

36. J. Powell, A. Henders, A. McRae, Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome …*, 456–466 (2012).

37. R. A. Irizarry *et al.*, The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–86 (2009).

38. A. Doi *et al.*, Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–3 (2009).

39. T. S. Mikkelsen *et al.*, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*. **448**, 553–60 (2007).

40. P. A. Futreal *et al.*, A census of human cancer genes. *Nat. Rev. Cancer*. **4**, 177–183 (2004).

41. C. E. Birdwell *et al.*, Genome-wide DNA methylation as an epigenetic consequence of Epstein-Barr virus infection of immortalized keratinocytes. *J. Virol.* **88**, 11442–58 (2014).

42. E. P. Brennan *et al.*, Comparative analysis of DNA methylation profiles in peripheral blood leukocytes versus lymphoblastoid cell lines. *Epigenetics*. **4**, 159–64 (2009).

43. Y. V. Sun *et al.*, Comparison of the DNA methylation profiles of human peripheral

blood cells and transformed B-lymphocytes. *Hum. Genet.* **127**, 651–658 (2010).

44. D. Grafodatskaya *et al.*, EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. *Genomics*. **95**, 73–83 (2010).

45. H. Sugawara *et al.*, Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines. *Epigenetics*. **6**, 509–516 (2011).

46. K. Åberg *et al.*, Methylome-wide comparison of human genomic DNA extracted from whole blood and from EBV-transformed lymphocyte cell lines. *Eur. J. Hum. Genet.* **20**, 953–5 (2012).

47. T. M. Thompson *et al.*, Comparison of whole-genome DNA methylation patterns in whole blood, saliva, and lymphoblastoid cell lines. *Behav. Genet.* **43**, 168–76 (2013).

48. M. Weber *et al.*, Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–66 (2007).

49. Y. Koga *et al.*, Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome Res.* **19**, 1462–1470 (2009).

Supplementary files

Supplementary Table 1: GO term enrichment analysis in all methylation sites.

Term (Biological Process)	Description	Count	%	*P*-Value	FDR
GO:0044707	single-multicellular organism process	1,343	38.1	2.31E-34	4.72E-31
GO:0007275	multicellular organism development	1,134	32.2	1.27E-32	2.59E-29
GO:0048731	system development	1,018	28.9	5.89E-32	1.20E-28
GO:0032501	multicellular organismal process	1,520	43.1	7.42E-30	1.51E-26
GO:0048856	anatomical structure development	1,230	34.9	4.43E-29	9.03E-26
GO:0044767	single-organism developmental process	1,230	34.9	5.27E-29	1.07E-25
GO:0032502	developmental process	1,258	35.7	9.93E-29	2.03E-25
GO:0030154	cell differentiation	865	24.5	7.31E-26	1.49E-22
GO:0048869	cellular developmental process	913	25.9	3.10E-22	6.32E-19
GO:0009653	anatomical structure morphogenesis	644	18.3	3.94E-22	8.03E-19
Term (Cellular Component)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0071944	cell periphery	1,251	35.5	4.22E-48	6.79E-45
GO:0005886	plasma membrane	1,218	34.6	3.31E-45	5.33E-42
GO:0044459	plasma membrane part	687	19.5	2.35E-35	3.79E-32
GO:0045202	synapse	259	7.3	7.60E-31	1.22E-27
GO:0097458	neuron part	388	11.0	6.31E-29	1.02E-25
GO:0044456	synapse part	202	5.7	2.32E-22	3.73E-19
GO:0005887	integral component of plasma membrane	429	12.2	5.64E-22	9.09E-19
GO:0031226	intrinsic component of plasma membrane	442	12.5	9.00E-22	1.45E-18
GO:0042995	cell projection	469	13.3	4.98E-21	8.03E-18
GO:0043005	neuron projection	279	7.9	2.68E-20	4.32E-17
Term (Molecular Function)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0022838	substrate-specific channel activity	150	4.3	3.88E-17	6.81E-14
GO:0022836	gated channel activity	122	3.5	4.65E-17	8.17E-14
GO:0046873	metal ion transmembrane transporter activity	144	4.1	5.33E-17	9.36E-14
GO:0005216	ion channel activity	145	4.1	2.14E-16	3.89E-13
GO:0005261	cation channel activity	113	3.2	4.01E-16	7.77E-13
GO:0022803	passive transmembrane transporter activity	154	4.4	2.61E-15	4.67E-12
GO:0015267	channel activity	153	4.3	4.93E-15	8.57E-12
GO:0005509	calcium ion binding	197	5.6	8.13E-12	1.43E-08
GO:0015075	ion transmembrane transporter activity	217	6.2	2.09E-11	3.67E-08
GO:0008092	cytoskeletal protein binding	223	6.3	2.37E-11	4.16E-08

Supplementary Table 2: GO term enrichment analysis in high-met-LCL.

Term (Biological Process)	Description	Count	%	*P*-Value	FDR
GO:0048518	positive regulation of biological process	160	39.2	6.32E-07	0.001201
GO:0048522	positive regulation of cellular process	147	36	6.94E-07	0.00132
GO:0006996	organelle organization	120	29.4	3.41E-06	0.006486
GO:0016043	cellular component organization	175	42.9	5.20E-06	0.009885
GO:0044237	cellular metabolic process	263	64.5	5.56E-06	0.010562
GO:0006366	transcription from RNA polymerase II promoter	68	16.7	1.04E-05	0.019721
GO:0030036	actin cytoskeleton organization	30	7.35	1.19E-05	0.022713
GO:0071840	cellular component organization or biogenesis	176	43.1	1.76E-05	0.033453
GO:0007010	cytoskeleton organization	48	11.8	1.87E-05	0.035525
GO:0051128	regulation of cellular component organization	79	19.4	4.58E-05	0.08703
Term (Cellular Component)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0031974	membrane-enclosed lumen	156	38.2	2.72E-12	4.05E-09
GO:0005622	intracellular	351	86	4.23E-12	6.29E-09
GO:0044424	intracellular part	345	84.6	6.87E-12	1.02E-08
GO:0043233	organelle lumen	153	37.5	9.14E-12	1.36E-08
GO:0070013	intracellular organelle lumen	151	37	1.08E-11	1.61E-08
GO:0044428	nuclear part	141	34.6	2.99E-11	4.45E-08
GO:0031981	nuclear lumen	131	32.1	3.30E-11	4.91E-08
GO:0043226	organelle	331	81.1	6.73E-11	1.00E-07
GO:0044446	intracellular organelle part	237	58.1	6.92E-11	1.03E-07
GO:0044422	organelle part	240	58.8	1.21E-10	1.79E-07
Term (Molecular Function)	**Description**	**Count**	**%**	***P*-Value**	**FDR**
GO:0005515	protein binding	289	70.8	1.19E-07	1.87E-04
GO:0001067	regulatory region nucleic acid binding	39	9.56	4.48E-05	0.070073
GO:0005488	binding	347	85	6.85E-05	0.107158
GO:0044212	transcription regulatory region DNA binding	38	9.31	8.93E-05	0.139794
GO:0016740	transferase activity	83	20.3	8.96E-05	0.140264
GO:0000975	regulatory region DNA binding	38	9.31	9.56E-05	0.14961
GO:0097159	organic cyclic compound binding	171	41.9	1.17E-04	0.183252
GO:1901363	heterocyclic compound binding	169	41.4	1.22E-04	0.191656
GO:1990837	sequence-specific double-stranded DNA binding	33	8.09	1.97E-04	0.308156
GO:0003690	double-stranded DNA binding	35	8.58	2.82E-04	0.441188

Appendix

R Script Code

Cluster analysis

```
## Data import ====
ClusterData   <-   read.table("cluster_30CpG.txt", header=T)

## Distance calculation (Euclidean distance) ====
d   =   dist(ClusterData)

## Cluster analysis with Ward method ====
hc_w   <-   hclust(d, method = "ward.D2")

## Save to png file ====
png("dendrogram.png", width=1860, height=492, unit="px")
plot(hc_w, cex = 0.5, ann = F, axes = F) #plot dendrogram
cluster <- rect.hclust(hc_w,k=2)
dev.off()

## Make a list of each cluster ====
write.table(cluster[1],file = "cluster1.txt", sep = "¥t") #First cluster
write.table(cluster[2],file = "cluster2.txt", sep = "¥t") #Second cluster
```

Making of the dataset for GO term enrichment analysis

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,20:23)]

## Making of the gene list containing logP>100 sites(All sites) ====
refIN <- subset(df,df$ref_IN == "IN")
logP100 <- subset(refIN,refIN$logP > 100)
a <- unique(logP100$ref_name)
A <- data.frame(ref = a)
write.table(A, "logP100Gene.txt", quote=F, sep="¥t", dec=".", row.names=F,
col.names=T)

## Making of the gene list containing logP>100 sites(high-met-LCL) ====
high_met <- subset(logP100, logP100$EB_PBL == "+")
h <- unique(high_met$ref_name)
H <- data.frame(ref = h)
write.table(H,   "high_logP100Gene.txt",   quote=F,   sep="¥t",   dec=".",
row.names=F, col.names=T)

## Making of the gene list containing logP>100 sites(low-met-LCL) ====
low_met <- subset(logP100, logP100$EB_PBL == "-")
l <- unique(low_met$ref_name)
L <- data.frame(ref = l)
write.table(L,   "low_logP100Gene.txt",   quote=F,   sep="¥t",   dec=".",
row.names=F, col.names=T)
```

The association with DNA methylation and Gene expression

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,20:23)]
df <- subset(df,df$ref_IN == "IN")
dfR <- nrow(df)
gewl <- read.table("gene expression wb vs lcl.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F, quote="")

## Add expression data to methylation sites ====
refgewl <- data.frame(NULL)
geR <- nrow(gewl)
  for(j in 1:geR){
    symbol_ge <- gewl[j, 7]
    ident_df <- subset(df,df$ref_name == symbol_ge)
    idR <- nrow(ident_df)
    if(idR != 0){
      for(k in 1:idR){
        ident_df$adj_P_val[k] <- gewl$adj.P.Val[j]
        ident_df$P_value[k] <- gewl$P.Value[j]
        ident_df$t[k] <- gewl$t[j]
        ident_df$B[k] <- gewl$B[j]
        ident_df$logFC[k] <- gewl$logFC[j]
      }
      refgewl <- rbind(refgewl,ident_df)
    }
  }

## Selection of the sites showing mehtylation level difference ====
refgewl_100 <- subset(refgewl,refgewl$logP > 100)

## Making graph in low-met-LCL and high-met-LCL ====
high_refgewl <- subset(refgewl_100,refgewl_100$EB_PBL == "+")
low_refgewl <- subset(refgewl_100,refgewl_100$EB_PBL == "-")
png("ref_high_Gene expression_wbvslcl.png", width = 450, height = 450)
plot(high_refgewl$logP, high_refgewl$logFC , pch=20,cex=0.5, axes=F,ann=F,
xlim=c(100,300), ylim=c(-5,5))
axis(1, pos=-5, xaxp=c(100,300,4), labels=F)
axis(2, pos=100, yaxp=c(-5,5,4), labels=F)
dev.off()
png("ref_low_Gene expression_wbvslcl.png", width = 450, height = 450)
plot(low_refgewl$logP, low_refgewl$logFC , pch=20,cex=0.5, axes=F,ann=F,
xlim=c(100,350), ylim=c(-8,8))
axis(1, pos=-8, xaxp=c(100,350,5), labels=F)
axis(2, pos=100, yaxp=c(-8,8,4), labels=F)
dev.off()
```

The association with DNA methylation and Gene expression in promoter region

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,17:19)]
```

```
df <- subset(df,df$tx_dis > -500 & df$tx_dis < 2000)
dfR <- nrow(df)
gewl <- read.table("gene expression wb vs lcl.txt",
                   header=TRUE, sep="¥t", stringsAsFactors=F, quote="")

## Add expression data to methylation sites ====
txgewl <- data.frame(NULL)
  geR <- nrow(gewl)
  for(j in 1:geR){
    symbol_ge <- gewl[j,7]
    ident_df <- subset(df,df$gene == symbol_ge)
    idR <- nrow(ident_df)
    if(idR != 0){
      for(k in 1:idR){
        ident_df$adj_P_val[k] <- gewl$adj.P.Val[j]
        ident_df$P_value[k] <- gewl$P.Value[j]
        ident_df$t[k] <- gewl$t[j]
        ident_df$B[k] <- gewl$B[j]
        ident_df$logFC[k] <- gewl$logFC[j]
      }
      txgewl <- rbind(txgewl,ident_df)
    }
  }

## Selection of the sites showing mehtylation level difference ====
txgewl_100 <- subset(txgewl,txgewl$logP > 100)

## Making graph in low-met-LCL and high-met-LCL ====
high_txgewl <- subset(txgewl_100,txgewl_100$EB_PBL == "+")
low_txgewl <- subset(txgewl_100,txgewl_100$EB_PBL == "-")
png("tx_high_Gene expression.png", width = 450, height = 450)
plot(high_txgewl$logP,  high_txgewl$logFC  ,  pch=20,cex=0.5,  axes=F,ann=F,
xlim=c(100,200), ylim=c(-5,5))
axis(1, pos=-5, xaxp=c(100,200,4), labels=F)
axis(2, pos=100, yaxp=c(-5,5,4), labels=F)
dev.off()
png("tx_low_Gene expression.png", width = 450, height = 450)
plot(low_txgewl$logP,  low_txgewl$logFC  ,  pch=20,cex=0.5,  axes=F,ann=F,
xlim=c(100,300), ylim=c(-7,7))
axis(1, pos=-7, xaxp=c(100,300,4), labels=F)
axis(2, pos=100, yaxp=c(-7,7,4), labels=F)
dev.off()
```

Volcano Plot in all sites

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,1:12]
df$PBL_EB <- df$ave_PBL12-df$ave_EB

## Save to png file ====
png("Volcano Plot_Global.png", width = 500, height = 400)
plot(df$PBL_EB, df$logP, type="n", ann=F, axes=F,
```

```
      xlim=c(-1,1), ylim=c(0,350))
axis(1, pos=0, xaxp=c(-1,1,4), labels=F)
axis(2, pos=0, yaxp=c(0,350,7), labels=F)
par(new=T)

for(i in 1:500){
   A    <- subset(df, ((i-1)/500*2)-1 < df$PBL_EB) #x-axis division(every
1/1000)
   B   <- subset(A, A$PBL_EB < (i/500*2)-1)
   for(j in 1:330){
      C   <- subset(B, (j-1) < B$logP) #y-axis division(every 1)
      D   <- subset(C, C$logP < j)
      r   <- nrow(D)
      plot(D$PBL_EB, D$logP, pch=20, cex=0.5,
         col = ifelse(r>100, "#FF0000FF",
               ifelse(r>80, "#FFFF00FF",
               ifelse(r>60, "#00FF00FF",
               ifelse(r>40, "#00FFFFFF",
               ifelse(r>20, "#0000FFFF",
               ifelse(r>10, "#FF00FFFF", "black")))))),
         ann = F, axes = FALSE, xlim = c(-1,1),
         ylim = c(0,350)) #Classifying with site density in each area
      par(new = T)
      }
}
dev.off()
```

Volcano Plot in each chromosome

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,1:12]
df$PBL_EB <- df$ave_PBL12-df$ave_EB

## Save to png file ====
png("Volcano Plot_EachCHR.png", width = 1600, height = 1300)
par(mfrow=c(4,6))
par(mar=c(2,2,2,2))

## plot ====
for (k in 1:22){
  df_chr <- subset(df,df$chr == k) #chromosome data
  plot(df$PBL_EB, df$logP, type="n", ann=F, axes=F,
       xlim=c(-1,1), ylim=c(0,350))
  axis(1, pos=0, xaxp=c(-1,1,4), labels=F)
  axis(2, pos=0, yaxp=c(0,350,2), labels=F)
  par(new=T)

  for(i in 1:500){
    A      <-  subset(df_chr,  ((i-1)/500*2)-1  <  df_chr$PBL_EB)  #x-axis
division(every 1/1000)
    B   <- subset(A, A$PBL_EB < (i/500*2)-1)
    for(j in 1:330){
      C   <- subset(B, (j-1) < B$logP) #y-axis division(every 1)
```

```
      D   <- subset(C, C$logP < j)
      r   <- nrow(D)
      plot(D$PBL_EB, D$logP, pch=20, cex=0.5,
           col = ifelse(r>100, "#FF0000FF",
                 ifelse(r>80, "#FFFF00FF",
                 ifelse(r>60, "#00FF00FF",
                 ifelse(r>40, "#00FFFFFF",
                 ifelse(r>20, "#0000FFFF",
                 ifelse(r>10, "#FF00FFFF", "black")))))),
           ann = F, axes = FALSE, xlim = c(-1,1),
           ylim = c(0,350)) # Classifying with site density in each area
      par(new = T)
    }
  }
  par(new = F)
}
dev.off()
```

TSS distance and P value in high-met-LCL

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,17,18)]

## Save to png file ====
png("txdis_high.png", width = 400, height = 500)

## Selecting the sites located in promoter region ====
prom <- subset(df,df$tx_dis>-1001&df$tx_dis<1001)
high <- subset(prom,prom$EB_PBL == "+")
plot(high$tx_dis, high$logP, type="n", ann=F, axes=F,
     xlim=c(-1000,1000), ylim=c(0,350))
par(new=T)

## Plotting the background data ====
plot(high$tx_dis, high$logP, pch = 20, cex = 0.5,
     xlim=c(-1000,1000),ylim=c(0,350),axes=F, ann=F,col="gray")
par(new=T)
axis(1, pos=0, xaxp=c(-1000,1000,10), labels=F)
axis(2, pos=-1000, yaxp=c(0,350,7), labels=F)
axis(4, pos=1000, yaxp=c(0,350,6), labels=F)
par(new=T)

## Calculating the proportion of differentially methylated sites ====
prom_ave <- subset(df,df$tx_dis>-1051&df$tx_dis<1051)
high_ave <- subset(prom_ave,prom_ave$EB_PBL == "+")
reg <- c(-1000:1000)
per_logP10 <- as.numeric(NULL)
per_logP25 <- as.numeric(NULL)
per_logP50 <- as.numeric(NULL)
for(i in 1:length(reg)){
  reg3 <- reg[i]-50
  reg5 <- reg[i]+50
  high_reg <- subset(high_ave,high_ave$tx_dis>reg3&high_ave$tx_dis<reg5)
  logP10 <- subset(high_reg,high_reg$logP > 10)
```

```
  logP25 <- subset(high_reg,high_reg$logP > 25)
  logP50 <- subset(high_reg,high_reg$logP > 50)
  per_logP10[i] <- nrow(logP10)/nrow(high_reg)
  per_logP25[i] <- nrow(logP25)/nrow(high_reg)
  per_logP50[i] <- nrow(logP50)/nrow(high_reg)
}
high_aveR <- data.frame(tx_dis=reg, logP10=per_logP10,
                        logP25=per_logP25, logP50=per_logP50)

## Plotting line graph of the sites proportion ====
plot(high_aveR$tx_dis,high_aveR$logP10,
      ylim = c(0,0.6),axes=F,ann=F,col="blue",pch=20, cex=0.5)
par(new=T)
plot(high_aveR$tx_dis,high_aveR$logP25,
     ylim = c(0,0.6),axes=F,ann=F,col="green",pch=20, cex=0.5)
par(new=T)
plot(high_aveR$tx_dis,high_aveR$logP50,
     ylim = c(0,0.6),axes=F,ann=F,col="red",pch=20, cex=0.5)
par(new=T)

dev.off()
```

TSS distance and P value in low-met-LCL

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,17,18)]

## Save to png file ====
png("txdis_low.png", width = 400, height = 500)

## Selecting the sites located in promoter region ====
prom <- subset(df,df$tx_dis>-1001&df$tx_dis<1001)
low <- subset(prom,prom$EB_PBL == "-")

plot(low$tx_dis, low$logP, type="n", ann=F, axes=F,
     xlim=c(-1000,1000), ylim=c(0,350))
par(new=T)

## Plotting the background data ====
plot(low$tx_dis, low$logP, pch = 20, cex = 0.5,
     xlim=c(-1000,1000),ylim=c(0,350),axes=F, ann=F,col="gray")
par(new=T)
axis(1, pos=0, xaxp=c(-1000,1000,10), labels=F)
axis(2, pos=-1000, yaxp=c(0,350,7), labels=F)
axis(4, pos=1000, yaxp=c(0,350,6), labels=F)
par(new=T)

## Calculating the proportion of differentially methylated sites ====
prom_ave <- subset(df,df$tx_dis>-1051&df$tx_dis<1051)
low_ave <- subset(prom_ave,prom_ave$EB_PBL == "-")

reg <- c(-1000:1000)
per_logP10 <- as.numeric(NULL)
per_logP25 <- as.numeric(NULL)
per_logP50 <- as.numeric(NULL)
for(i in 1:length(reg)){
  reg3 <- reg[i]-50
  reg5 <- reg[i]+50
  low_reg <- subset(low_ave,low_ave$tx_dis>reg3&low_ave$tx_dis<reg5)
  logP10 <- subset(low_reg,low_reg$logP > 10)
  logP25 <- subset(low_reg,low_reg$logP > 25)
  logP50 <- subset(low_reg,low_reg$logP > 50)
  per_logP10[i] <- nrow(logP10)/nrow(low_reg)
  per_logP25[i] <- nrow(logP25)/nrow(low_reg)
  per_logP50[i] <- nrow(logP50)/nrow(low_reg)
}
low_aveR <- data.frame(tx_dis=reg, logP10=per_logP10,
                       logP25=per_logP25, logP50=per_logP50)

## Plotting line graph of the sites proportion ====
plot(low_aveR$tx_dis,low_aveR$logP10,
     ylim = c(0,0.6),axes=F,ann=F,col="blue",pch=20, cex=0.5)
par(new=T)
plot(low_aveR$tx_dis,low_aveR$logP25,
```

```
       ylim = c(0,0.6),axes=F,ann=F,col="green",pch=20, cex=0.5)
par(new=T)
plot(low_aveR$tx_dis,low_aveR$logP50,
     ylim = c(0,0.6),axes=F,ann=F,col="red",pch=20, cex=0.5)
par(new=F)

dev.off()
```

The association with TSS distance and CpG island

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:18)]

## Selecting the sites located in promoter region ====
prom_ave <- subset(df,df$tx_dis>-1051&df$tx_dis<1051)

## Classifying with high-met or low-met ====
high <- subset(prom_ave,prom_ave$EB_PBL == "+")
low <- subset(prom_ave,prom_ave$EB_PBL == "-")

## Classifying with in CpG island or out ====
high_IN <- subset(high,high$CpG_island == "IN")
high_OUT <- subset(high,high$CpG_island == "OUT")
low_IN <- subset(low,low$CpG_island == "IN")
low_OUT <- subset(low, low$CpG_island == "OUT")

## Save to png file ====
png("CGI_txdis.png", width=700, height=600)

## Calculating and plotting the proportion of the logP>10 sites ====
CGItx_list <- list(high_IN, high_OUT, low_IN, low_OUT)
list_col <- c("sky blue", "blue", "orange", "red")
for(j in 1:4){
  DF <- as.data.frame(CGItx_list[j])
  reg <- c(-1000:1000)
  per_logP10 <- as.numeric(NULL)
  for(i in 1:length(reg)){
    reg3 <- reg[i]-50
    reg5 <- reg[i]+50
    DF_reg <- subset(DF,DF$tx_dis>reg3 & DF$tx_dis<reg5)
    logP10 <- subset(DF_reg,DF_reg$logP > 10)
    per_logP10[i] <- nrow(logP10)/nrow(DF_reg)
  }
  DF_aveR <- data.frame(tx_dis=reg,logP10=per_logP10)
  plot(DF_aveR$tx_dis,DF_aveR$logP10,
       ylim = c(0,0.8),axes=F,ann=F,col=list_col[j],pch=20, cex=0.5)
  par(new=T)
}
axis(1, pos=0, xaxp=c(-1000,1000,10), labels=F)
axis(2, pos=-1000, yaxp=c(0,0.8,4), labels=F)
dev.off()
```

The methylation level distribution in HCPs, ICPs, and LCPs

```
## Data import ====
df <- read.table("selection_P_CGI_refG_GIO_HIL_txdis.txt",
                 header=TRUE, sep="¥t", stringsAsFactors=F)
df <- df[,c(1:12,24,25)]

## Classifying with HCP, ICP, or LCP ====
HCP <- subset(df,df$HIL == "HCP")
ICP <- subset(df,df$HIL == "ICP")
LCP <- subset(df,df$HIL == "LCP")
HIL <- list(HCP,ICP,LCP)

## Calculating the proportion of the sites in every methylation level (PBC)
====
PBC_freq <- NULL
for(i in 1:3){
  A <- as.data.frame(HIL[i])
  met <- seq(0,1, by=0.125)
  num <- as.numeric(NULL)
  for(j in 1:length(met)-1){
    A_reg <- subset(A,A$ave_PBL12>met[j] & A$ave_PBL12<met[j+1])
    num_reg <- nrow(A_reg)
    num[j] <- num_reg
  }
  num_freq <- num/nrow(A)*100
  PBC_freq <- cbind(PBC_freq,num_freq)
}
write.table(PBC_freq,   "PBC_HIL_freq.txt",   quote=F,   sep="¥t",   dec=".",
row.names=F, col.names=T)

## Calculating the proportion of the sites in every methylation level (LCL)
====
LCL_freq <- NULL
for(i in 1:3){
  A <- as.data.frame(HIL[i])
  met <- seq(0,1, by=0.125)
  num <- as.numeric(NULL)
  for(j in 1:length(met)-1){
    A_reg <- subset(A,A$ave_EB>met[j] & A$ave_EB<met[j+1])
    num_reg <- nrow(A_reg)
    num[j] <- num_reg
  }
  num_freq <- num/nrow(A)*100
  LCL_freq <- cbind(LCL_freq,num_freq)
}
write.table(LCL_freq,   "LCL_HIL_freq.txt",   quote=F,   sep="¥t",   dec=".",
row.names=F, col.names=T)

## Checking the alteration in every methylation level ====
pName_HCP <- c("HCP_000-025.png","HCP_025-050.png",
               "HCP_050-075.png","HCP_075-100.png")
pName_ICP <- c("ICP_000-025.png","ICP_025-050.png",
                "ICP_050-075.png","ICP_075-100.png")
```

```
pName_LCP <- c("LCP_000-025.png","LCP_025-050.png",
               "LCP_050-075.png","LCP_075-100.png")
plot_name <- list(pName_HCP,pName_ICP,pName_LCP)
met <- seq(0,1, by=0.25)
for(i in 1:3){
  A <- as.data.frame(HIL[i])
  A$PBL_EB <- A$ave_PBL12 - A$ave_EB
  pName <- as.data.frame(plot_name[i])
  for(j in 1:4){
    Amet <- subset(A,A$ave_PBL12 > met[j] & A$ave_PBL12 < met[j+1])
    png(pName[j,],width = 150,height = 100)
    par(mar=c(0,0,0,0))
    plot(Amet$PBL_EB, Amet$logP, xlim=c(-1,1), ylim=c(0,300),ann=F, axes=F,
pch=20, cex=0.5)
    axis(1, pos=0, xaxp=c(-1,1,4), labels=F)
    axis(2, pos=0, yaxp=c(0,300,2), labels=F)
    dev.off()
  }
}
```