Studies into epigenetic variation and its contribution to cardiovascular disease

Talens, R.P.

Chapter 6

Genome Wide Differences and Similarities in DNA Methylation Between Internal Tissues and Blood

Rudolf P Talens,[1] Steffan D Bos,[1,2] Maurits RA Drijfhout van Hooff,[1] Judith VMG Bovée,[3] Ruud van der Breggen,[1] Nico Lakenberg,[1] Erik B van den Akker,[1,4] J Wouter Jukema,[2,5] P Eline Slagboom,[1,2] Ingrid Meulenbelt,[1,2] Jelle J. Goeman,[6] Bastiaan T. Heijmans[1,2]

1. Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
2. Netherlands Consortium for Healthy Ageing
3. Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands
4. Department of Mediamatics, Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
5. Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands
6. Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

Data presented in this manuscript was augmented with additional data, reanalysed and published in Slieker et al., Epigenetics & Chromatin, Volume 6, article 26, pages 1 – 12, 2013.

Abstract

Background
Human studies on epigenetic variation and disease are usually restricted to measuring peripheral tissues, while many diseases manifest in internal tissues. Due to the extensive epigenetic remodeling involved in tissue development, it is unclear to what extent inter-individual epigenetic variation measured in peripheral tissues marks that of the internal tissues related to the disease.

Methodology & Principal Findings
To investigate epigenetic differences and similarities between blood and internal tissues we obtained samples from 6 individuals during post-mortem examination within 24 hours after death. Using the Illumina Infinium HumanMethylation450 BeadChip microarray We compared DNA methylation of blood with subcutaneous fat, skeletal muscle, visceral fat, liver, pancreas and spleen at 378,239 CpG sites distributed over the autosomal chromosomes. Principal component analysis on this data revealed that most differences in genome wide DNA methylation patterns were between tissues rather than individuals. This analysis also revealed that the methylation patterns in post-mortem blood samples were similar to the patterns in blood samples form 40 middle aged regular blood donors. The tissues showed a similar overall distribution of mean methylation levels and methylation variation at individual CpG sites. A quarter of the CpG sites measured showed no detectable methylation difference between blood and any of the internal tissues. Further, the methylation variation at more than 30% of CpG sites showed a strong correlation ($r \geq 0.75$) between blood and an internal tissue.
Most strong correlations were between blood and one internal tissue exclusively, although strong correlations between blood and several tissues were observed at 6092 CpG sites. searching for characteristics marking CpG sites with a strong tissue correlation, we found that these were unrelated to DNA methylation levels, variation, or differences between the tissues, nor did we observe an association with genetic sequence motifs, genome structure, or localization relative to genomic, epigenomic, or repetitive elements.

Conclusions

Variation in DNA methylation is correlated between blood and internal tissues at a subset of CpG sites, depending on both the genomic locus and the tissue. These results indicate the existence of loci at which epigenetic information in peripheral tissues marks that of internal tissues.

Introduction

A popular theme in research on the development of common age related diseases is its epigenetic component [17,65,68,231]. Epigenetic mechanisms regulate the local condensation of DNA thereby influencing the capacity of the transcription machinery to access a locus upon reception of the correct cues [21,23,29]. There are several correlated layers of epigenetic information, including post-transcriptional control by microRNAs, nucleosomal packaging, histone modifications and DNA methylation [22–25]. Epigenetic remodeling of cellular expression potential during embryogenesis is commonly thought to create the different cellular phenotypes from the single genotype of a zygote [182]. Epidemiological studies investigating the epigenetic component of common diseases mainly focus on DNA methylation since it can be reliably measured on DNA samples that are stored in a biobank, even after several decades [195].

Associations between locus specific DNA methylation and risk for rheumatoid arthritis [232], obesity [48] and myocardial infarction [233] have recently been reported. However, as bone [232], visceral fat [48] and myocardium [233] are unavailable for such studies, these associations were found in blood, which is readily available but not directly involved in the disease. Thus, the epigenetic contribution to these diseases remains unclear, since the relation between DNA methylation in blood and that of other tissues is not yet established. This is a pressing issue in epigenetic epidemiological research for most diseases. Clinical biobanks store DNA that was extracted from peripheral tissues, usually blood, occasionally buccal cells, and sporadically skin, subcutaneous fat, and skeletal muscle. This requires epigenetic studies to carefully consider the availability of tissues and the relevance of the information that can be obtained [98,99,202].

Documenting the loci at which DNA methylation measured in blood is a useful marker for inaccessible disease related tissues, will benefit interpreting the results of epigenetic studies, and may help even designing these studies if many such loci exist. DNA methylation differences between tissues have been reported at many loci [100–102,234]. Importantly though, despite the focus of these studies on tissue differences, they also observed genomic areas with similar DNA methylation between the tissues within an individual [100,234]. Moreover, a study on candidate loci found a strong correlation between DNA methylation in blood and buccal swabs at 4 of the 8 loci investigated [195] and a genome wide study found loci with correlations between blood and brain cortex or cerebellum [234]. Together these results suggest that for a potentially subset of loci, DNA methylation in blood may mark the methylation status of another (disease related) tissue.

In this study we investigate epigenetic similarities between blood and several internal tissues. We obtained post mortem samples of blood, subcutaneous fat (SC fat), muscle, visceral fat (VS fat), liver, spleen, and pancreas from six individuals. We use the Illumina Infinium HumanMethylation450 BeadChip microarray (Illumina, San Diego, USA) to measure DNA methylation at 485,462 CpG sites distributed throughout the genome [85]. We focused on the 378,239 CpG sites for which the measurement was not influenced by known genetic variation. Principal component analysis (PCA) was used to investigate genome wide patterns of DNA methylation. At each of these 378,239 CpG sites we compared average DNA methylation and computed correlation coefficients between tissues. Our exploration of tissue correlations focused on 219,558 CpG sites which had multiple nearby CpG sites measured, using smoothed correlations over these CpG sites to indicate a more robust local effect [195].

Methods and materials

Subjects

The samples in this study were collected during post-mortem examination (within 24 hours after death) of 6 individuals (3 men) between 58 and 79 years old. From each individual we collected samples of 6 different tissues. Samples of blood, skeletal muscle, subcutaneous fat (SC fat) and visceral fat (VS fat) were collected from each individual, liver samples were collected from 5 individuals, pancreas from 4 individuals and spleen from 3 individuals. Autopsy case with tissue sample information is listed in Table 6.1. Blood was drawn from the carotid arteries, collected in BD

Table 6.1: Characteristics of selected individuals

Subject#	Gender	Age	PMI (h)	Cause of death	Tissues						
1	Female	61	7	Sepsis	B	F	V	M	L	S	
2	Male	58	12.5	Heart failure	B	F	V	M	L	S	
3	Male	64	9	Myocardial infarction	B	F	V	M	L		P
4	Female	65	15	Aortic aneurysm	B	F	V	M	L		P
5	Female	66	10	Liver cirrhosis	B	F	V	M		S	P
6	Male	79	10	Sepsis	B	F	V	M	L		P

PMI: Post Mortem Interval in hours

Tissues: B = Blood, F = Subcutaneous fat, V = Visceral fat, M = Muscle, L = Liver, S = Spleen, P = Pancreas

vacutainer EDTA blood collection tubes (Becton, Dickinson and Company, Franklin Lakes, NJ, USA), and treated according to standard protocol. The other tissues collected were dissected at the time of autopsy, glued to a piece of cork using Tissue-Tek OCT Compound (Sakura Finetek, Alphen aan de Rijn, Netherlands), snap-frozen in liquid nitrogen and stored at -80° C until further use. With a CM3000 Cryostat (Leica, Wetzlar, Germany) cryosections were made of the tissues. Two slices of 5 µm thick at either end of a tissue sample were stained with hematoxylin and eosin (H&E) to check tissue integrity and cell type heterogeneity, since each sample represent a mixture of the cell types normally found in that organ. All tissue from which DNA was isolated were histologically normal (Supplementary figure S6.1) and of similar consistency between the two HE-stained section of an individual, indicating an even heterogeneity throughout the tissue sample. Tissue heterogeneity was also similar between individuals, except for visceral fat where substantial differences in heterogeneity were observed between individuals (Supplementary figure S6.1).

DNA from blood was extracted from the buffy coat layer after centrifugation. Tissue in between the histological slides was cut into 20 µm thick slices, 30 to 40 of which were lysed overnight at 56° C in QIAGEN ATL buffer + proteinase K (Qiagen, Düsseldorf, Germany). DNA was extracted using a double Phenol : Chloroform : Isoamyl alcohol (25:24:1) (PCI) separation. The supernatant of the first separation, was treated with PCI once again to remove all residual traces of fatty acids, which was especially necessary for tissues with a high lipid content. The remainder of the procedure followed the standard protocol.

All samples were obtained with approval of the pathology department of the LUMC and according to the ‘Proper Secondary Use of Human Tissue’ code of conduct by the ‘Federatie van Medisch Wetenschappelijke Verenigingen’.

Genome wide DNA methylation

DNA methylation was measured at 485,462 CpG sites using the Illumina Infinium HumanMethylation450 BeadChip array (Illumina, San Diego, USA) according to manufacturer’s protocol [85]. In short: about 700 ng of genomic DNA was bisulphite converted using the EZ-96 DNA Methylation Kit (Zymo Research, Orange County, USA). Each DNA sample was whole-genome amplified, and enzymatically fragmented. The DNA fragments were hybridized to locus-specific DNA oligomers linked to individual bead types. After single base pair extension, the array was fluorescently stained, scanned and the intensities of the non-methylated and methylated bead types were measured.

All samples of an individual were measured on the same chip. To check whether DNA methylation is affected by death, we compared DNA methylation of the post mortem blood samples with DNA from blood samples of living blood donors (35 – 40 years old), who represent DNA methylation in the healthy adult population [235]. This DNA was measured in two mixtures, one containing DNA of twenty men, the other of twenty women. Methylation of pooled DNA samples was recently demonstrated to accurately reflect the group average [236].

Quality control

Technical variability between chips was tested by measuring a duplo series of methylation titers from standardized non-methylated and methylated DNA samples (0%, 25%, 50%, 75%, 100% methylated; Zymo Research, Orange County, USA). Between chip variability was found negligible (Supplementary figure S6.2). All samples met the criteria tested with a fixed set of technical probes for bisulphite conversion, background signal, and hybridization efficiency, according to the Controls Dashboard from Illumina’s GenomeStudio software [85]. Probe performance within each sample was investigated with the detection

P-value. It compares the signal generated at each CpG site to negative controls and can be interpreted as the probability of seeing a certain signal level without specific probe-target hybridization. Probes with detection P-value ≥ 0.05 in one of the samples were discarded for all samples before normalization with the Simple Scaling Normalization (SSN) method of the R-package *lumi* [237], which relies on signal intensity and methylation patterns (Supplementary figure S6.3).

In this study we focused on probes assessing CpG methylation on autosomal chromosomes, discarding the 11,650 probes on X- and Y- chromosomes. As annotated by the manufacturer, 86,769 contain a SNP (dbSNP 30) in their 50 bp target sequence or the CpG dinucleotide. Since it was shown on a earlier version of the platform that this could interfere with methylation measurements [238], we discarded measurements at these CpG site for further analyses. In all, we analyzed methylation measurements at 378,239 CpG sites in all six tissues for all six individuals.

Analyses

Comparing DNA methylation between the tissues

DNA methylation results from the Beadchip array platform can be described by two inter-exchangeable values, the ß-value and the M-value [239]. The ß-value represents DNA methylation on an interpretative scale of 0 to 1. The M-value represents it on a quantitatively more accurate logarithmic (log2) scale that resolves the skewed distribution in the upper (ß > 0.8; M > 2) and lower (ß < 0.2; M < -2) methylation levels [239]. For analyses and calculations the M-value was used for methylation levels of CpG sites, for presentation purposes the relevant results were transformed into the ß-value.

For genome wide DNA methylation patterns, principal component analysis (PCA) was used to investigate inter-individual variation vs. between-tissue (intra-individual) variation. At individual CpG sites, methylation differences between blood and the other tissues were tested with a two-sided t-test, and mean methylation, transformed into ß-value, and methylation variation, represented with the size of the standard deviation (SD) interval (mean ± 1*SD) in ß-value, were compared. Further, relative similarities between blood and the tissues were investigated using the Pearson correlation coefficient (r). From here on in this study, tissue correlation, unless specified, refers to the correlation of DNA methylation in blood with that of another tissue, and tissue difference, unless specified, refers to differences in mean DNA methylation between blood and another tissue.

Smoothing tissue correlations and tissue differences

Previous research has demonstrated that methylation of nearby CpG sites is often correlated [195] and that this intra-class correlation can be used as a powerful tool to reveal subtle differences in DNA methylation that are consistently similar across several nearby CpG sites [135,233]. Exploiting this principle, we applied a smoothing algorithm with a sliding bandwidth of 2,000 bases centered around the cytosine residue of the CpG site under investigation. This created a more robust measure for tissue correlations and differences at individual CpG sites, partly compensating for the small amount of individuals in this study. After smoothing the absolute difference was used in combination with results from the two-sided t-test to describe tissue differences. For tissue correlations we discriminated between CpG sites with fewer than 3 and those with 3 or more additional CpG sites within the bandwidth (Supplementary figure S6.4). The smoothed correlation at the latter CpG sites was considered indicative of a more robust local effect, representing the mean over 4 or more CpG sites. We focussed our exploration of tissue correlations on these CpG sites, which creates a bias for more CpG rich genomic areas, as CpG poor areas are less likely to meet these criteria. A smoothed correlation was considered strong when $r > 0.75$. Note that in this study all negative tissue correlations, even when strong ($r < -0.75$), were given the same weight as weak positive correlations, since, although mathematically equal to positive correlations, there is to our knowledge no experimental data in support of a biological interpretation of negative tissue correlation.

To scan for genomic regions enriched in tissue correlations we divided the genome in blocks with a length of 100,00 bp. For each block, we counted the number of probes with at least 3 additional CpG sites within the smoothing bandwidth (CpG rich area probes), the number of such probes with a strong tissue correlation, the proportion of correlated probes (for blocks ≥ 15 probes counted), the number of probes with strong tissue correlation between blood and 1 or 2 tissues, between blood and 3 to 5 tissues, and between blood and all 6 tissues. We plotted this information against the human genome (assembly hg18/NCBI36) using the genome graph function of the UCSC genome browser [150] (website: http://genome.ucsc.edu/).

Identifying characteristics of tissue correlations

To investigate whether CpG sites with a strong correlation can be identified by characteristics of DNA methylation, we first explored the relation of tissue correlation with mean methylation, methylation variation, and tissue differences. For this means and SD interval sizes were categorized based on their observed distributions across all CpG sites, and tissue differences based on previously published observations on inter individual and between tissue differences [100,195,233]. Mean methylation categories: low (ß < 0.2), below medium (0.2 ≤ ß < 0.4), medium (0.4 ≤ ß < 0.6), above medium (0.6 ≤ ß < 0.8), and high (ß ≥ 0.8). Methylation variation categories: low (SD interval < 0.05), intermediate (0.05 ≤ SD interval < 0.15), and high (SD interval ≥ 0.15). Tissue difference categories: high (absolute difference ≥ 0.2 & $p_{t\text{-test}} \leq 0.05$), intermediate (0.05 ≤ absolute difference ≥ 0.2 & $p_{t\text{-test}} \leq 0.05$), low (absolute difference ≤ 0.05 & $p_{t\text{-test}} \leq 0.05$), and insignificant ($p_{t\text{-test}} > 0.05$). The distributions across these categories were inspected for CpG sites from CpG poor areas, CpG sites from CpG rich areas with no strong tissue correlation and CpG sites from CpG rich areas with a strong correlation with at least one tissue. Note that a previous version of the platform already observed that CpG rich areas have different distributions of mean methylation and methylation variation compared with CpG poor areas [90].

To investigate whether CpG sites with a strong correlation can be identified by the function of their genomic area as annotated by the manufacturer [85] (Supplementary table S6.2, for full list of genomic locations) we inspected frequency distributions of these locations between strong correlations with 3 or more tissues, strong correlations with 1 or 2 tissues and weak correlations with all tissues in CpG rich areas, and also between strong and weak correlations in CpG poor areas.

To test for structural differences, CpG sites with the strongest correlations (r > 0.75 for at least 1 tissue correlation, and r > 0.25 for all tissue correlations; 6,539

CpG sites) were compared to CpG sites with the weakest correlations ($r < 0.25$ for all tissue correlations; 5,837 CpG sites), all located in CpG rich areas. A Chi^2 test was used to discern differences in the distributions of their genomic location (as annotated by the manufacturer). The Wilcoxon test of the epigraph web tool [240] (website: http://epigraph.mpi-inf.mpg.de/WebGRAPH/) was used to investigate differences in immediately adjacent DNA sequence (base composition and 2-mers, +/- 10 bp from C residue), location in transcription factor binding sites (TFBS) and repeats, and all available data on DNA structure, chromosome organization, and epigenome and chromatin structure (histone code in blood) [240].

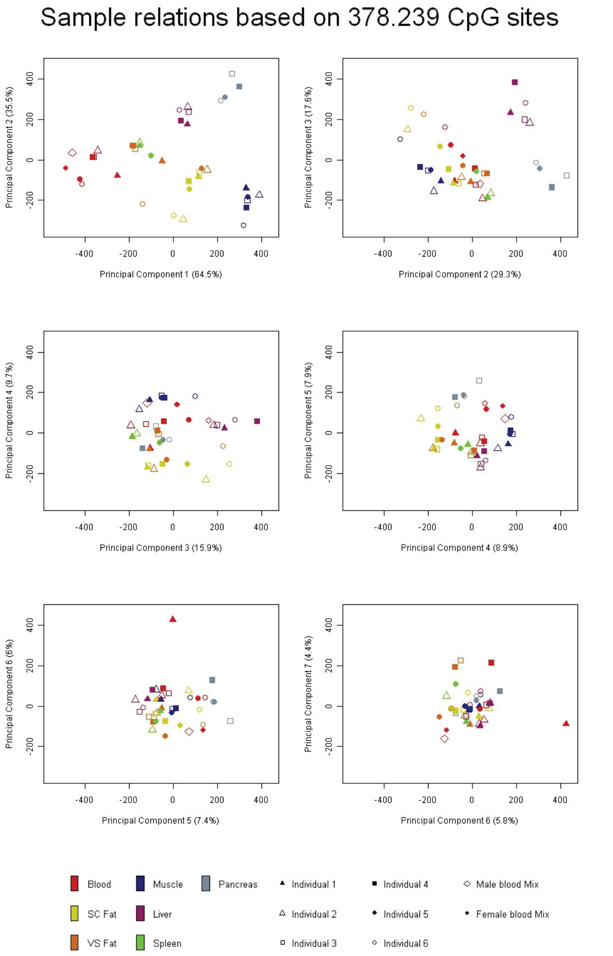


Figure 6.1: Scatter plots of principal components 1 to 7, derived from the Principal Component Analysis comparing the genome wide DNA methylation patterns of the tissue samples. In the top left corner the two lowest components (1st and 2nd) are plotted against each other, in the bottom right the two highest (6th and 7th). The lowest of two components is on the x-axis, the percentage giving on the axes is a measure of the amount variation represented by the component in this analyses. The symbols plotted are specific for samples of each individual, with open symbols for men, and a color specific to samples of each tissue. The distance between the samples is a measure of their similarity.

Results

PCA analysis on genome wide methylation patterns

A genome wide survey resulted in DNA methylation data across seven tissues of six individuals at 378,239 CpG sites distributed throughout the genome. Variation of genome wide patterns in DNA methylation was investigated by visualizing results from principal component analysis (PCA; Figure 6.1). Plotting the 1st vs. the 2nd principal components, covering almost all variation between the genome wide methylation patterns, showed tight clusters of the samples from the same tissue of each individuals (Figure 6.1, top left), even plotting the 6th vs. 7th principal component, covering roughly 10 % of variation, still showed no clusters of the samples from the various tissues of the same individual (Figure 6.1, bottom right). In this analysis the male and female blood mixes, representing healthy individuals, were observed to cluster together with the post mortem blood samples up to the 5th principal component, in the 6th and 7th components the blood mixes separated somewhat from the rest of the blood samples, but from the fourth component onward the analysis revealed no specific clusters of DNA samples (Figure 6.1).

Mean methylation and variation per tissue

The distribution of mean methylation across all 378,239 CpG sites investigated showed a similar bimodal shape in all tissues with peaks at low and high methylation and a valley at medium methylation (Figure 6.2). The distribution of methylation variation across all CpG sites also had a skewed shape similar for each tissue, with the SD interval peaking around 0.05 in ß-value and a long tail of higher variation, starting at SD interval around 0.2 in ß-value, and continuing to a maximum SD interval around 0.9 in ß-value (Figure 6.3).

Inspecting tissue differences at individual CpG sites revealed that almost 74 % of CpG sites showed a nominally significant ($p_{\text{T-test}} < 0.05$) difference in DNA methylation between blood and at least one of the other tissues (Table 6.2). Looking at the comparison of blood with each tissue, at the most 48 % of CpG sites, in the case of the spleen, and at the least 9 % of CpG sites, in the case of the pancreas, were differently methylated ($p_{\text{T-test}} < 0.05$). Most CpG sites with tissue differences had a different DNA methylation between blood and multiple tissues. For the spleen the biggest set of

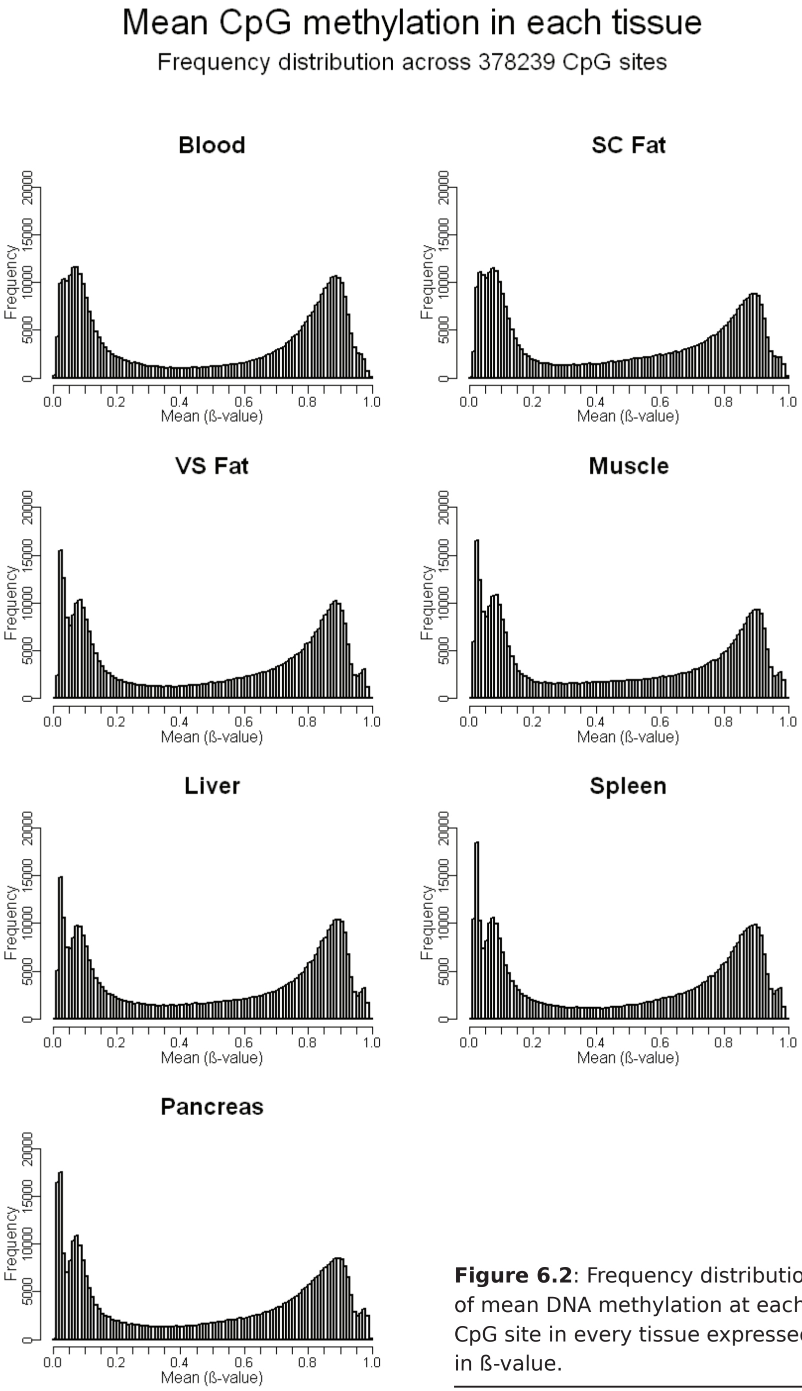


Figure 6.2: Frequency distributions of mean DNA methylation at each CpG site in every tissue expressed in ß-value.

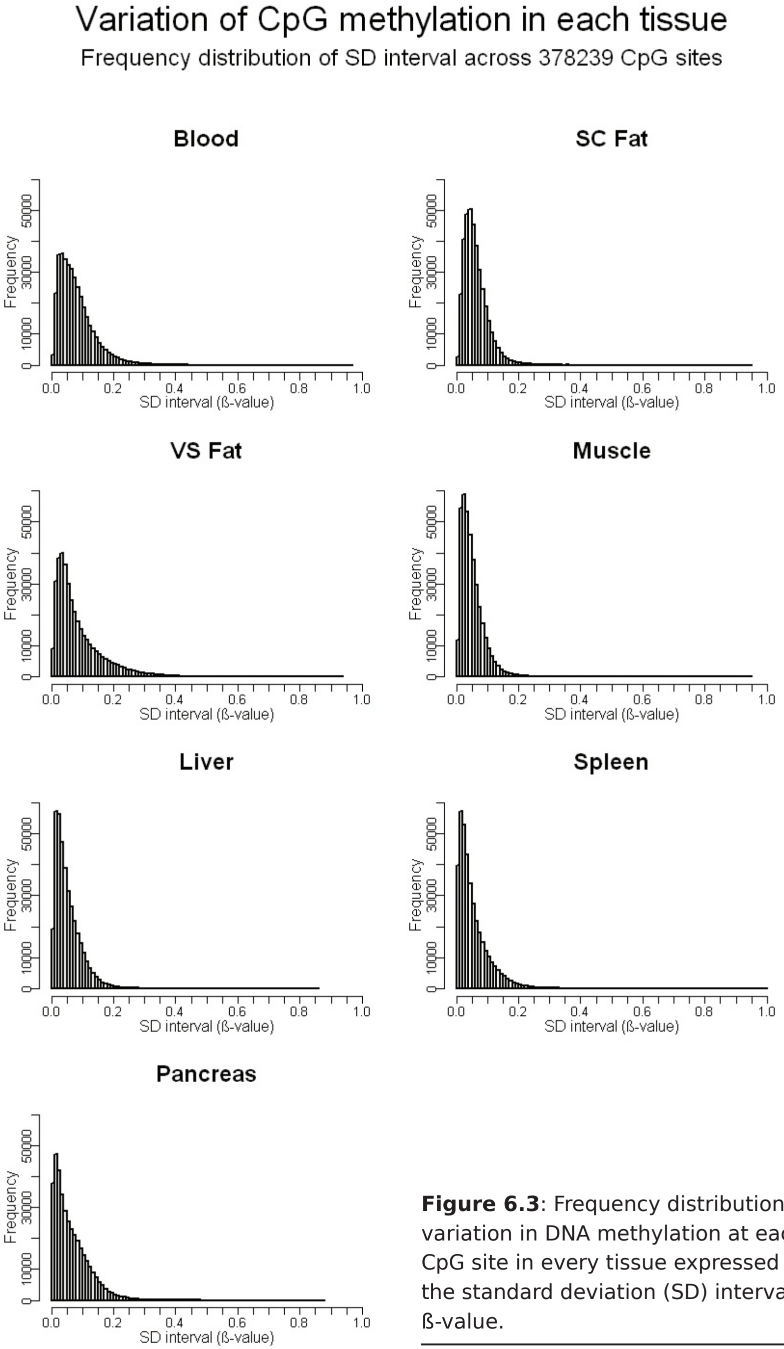


Figure 6.3: Frequency distributions of variation in DNA methylation at each CpG site in every tissue expressed as the standard deviation (SD) interval in ß-value.

Table 6.2: Significant methylation differences between blood and other tissues across 378,239 CpG sites

Amount of tissues[a]	Amount (percentage) of CpG sites with significant ($P_{T\text{-test}} < 0.05$) methylation differences													
	SC Fat		VS Fat		Muscle		Liver		Spleen		Pancreas		All tissues	
0	255426	(67.5 %)	295548	(78.1 %)	204087	(54.0 %)	247602	(65.5 %)	197537	(52.2 %)	344186	(91.0 %)	98618	(26.1 %)
1	5573	(1.5 %)	2533	(0.7 %)	18786	(5.0 %)	14357	(3.8 %)	44549	(11.8 %)	2619	(0.7 %)	88417	(23.4 %)
2	18005	(4.8 %)	6705	(1.8 %)	40505	(10.7 %)	24861	(6.6 %)	36786	(9.7 %)	3554	(0.9 %)	65208	(17.2 %)
3	30887	(8.2 %)	14427	(3.8 %)	42488	(11.2 %)	27055	(7.2 %)	33250	(8.8 %)	3723	(1.0 %)	50610	(13.4 %)
4	30201	(8.0 %)	20494	(5.4 %)	33546	(8.9 %)	26573	(7.0 %)	28530	(7.5 %)	4800	(1.3 %)	36036	(9.5 %)
5	24656	(6.5 %)	25041	(6.6 %)	25336	(6.7 %)	24300	(6.4 %)	24096	(6.4 %)	5866	(1.6 %)	25859	(6.8 %)
6	13491	(3.6 %)	13491	(3.6 %)	13491	(3.6 %)	13491	(3.6 %)	13491	(3.6 %)	13491	(3.6 %)	13491	(3.6 %)
Total[b]	122813	(32.5 %)	82691	(21.9 %)	174152	(46.0 %)	130637	(34.5 %)	180702	(47.8 %)	34053	(9.0 %)	279621	(73.9 %)

a: The number tissues with which these CpG sites have a different methylation compared with blood

b: The total amount of CpG sites that have a different DNA methylation between blood and this tissue differently methylated CpG sites were uniquely so between blood and spleen, whereas for the pancreas the biggest set was differently methylated between blood and all six tissues (Table 6.2).

Correlations of CpG methylation between tissues

An alternative way to express tissue similarity, which takes into account differences in inter- and intra-individual (between-tissue) variation in DNA methylation across the tissues, is to compute correlation coefficients between blood and the other tissues. To inspect tissue correlations at a local genomic scale the correlation coefficients of each CpG site were smoothed over all measured CpG sites within the bandwidth of 2000 bp. For 92,916 CpG sites there were no additional CpG sites within the 2000 bp bandwidth (Supplementary figure S6.4). The other 285,323 CpG sites had from 1 up to 59 additional CpG sites within the bandwidth. The smoothed correlation was considered indicative of a local genomic effect when the bandwidth contained at least 4 CpG sites. The resulting smoothed correlation of such CpG sites thus represents a mean of measurements at 4 to 60 CpG sites. To illustrate this, the methylation values of blood against the other tissues were plotted for each of the 5 CpG sites within the bandwidth around probe cg2201694 (Figure 6.4).

Inspecting the smoothed correlations revealed that over 30 % of CpG sites had a strong correlation ($r > 0.75$) between blood and at least one other tissue. This was true for CpG rich (219,558 CpG sites) and CpG poor (158,681 CpG sites) areas (Table 6.3A). Looking at the distributions of tissue correlations for each tissue at the most 15.5 %, in the case of the spleen, and at the least 3.2 %, in the case of muscle, showed a strong correlation with blood. Remarkably, the endodermal tissues had more CpG sites strongly correlated with blood than the mesodermal tissues. Liver, the endodermal tissue with the lowest amount of CpG sites (15,444 sites) with a strong tissue correlation, still had almost 2,000 such CpG sites than SC fat, which was the mesodermal tissue with the highest amount CpG sites (13,501 sites) with a strong tissue correlation (Table 6.3A).

Looking in CpG rich areas at CpG sites with a strong correlation between blood and more than one tissue, there were 3 times more CpGs exclusively correlated between

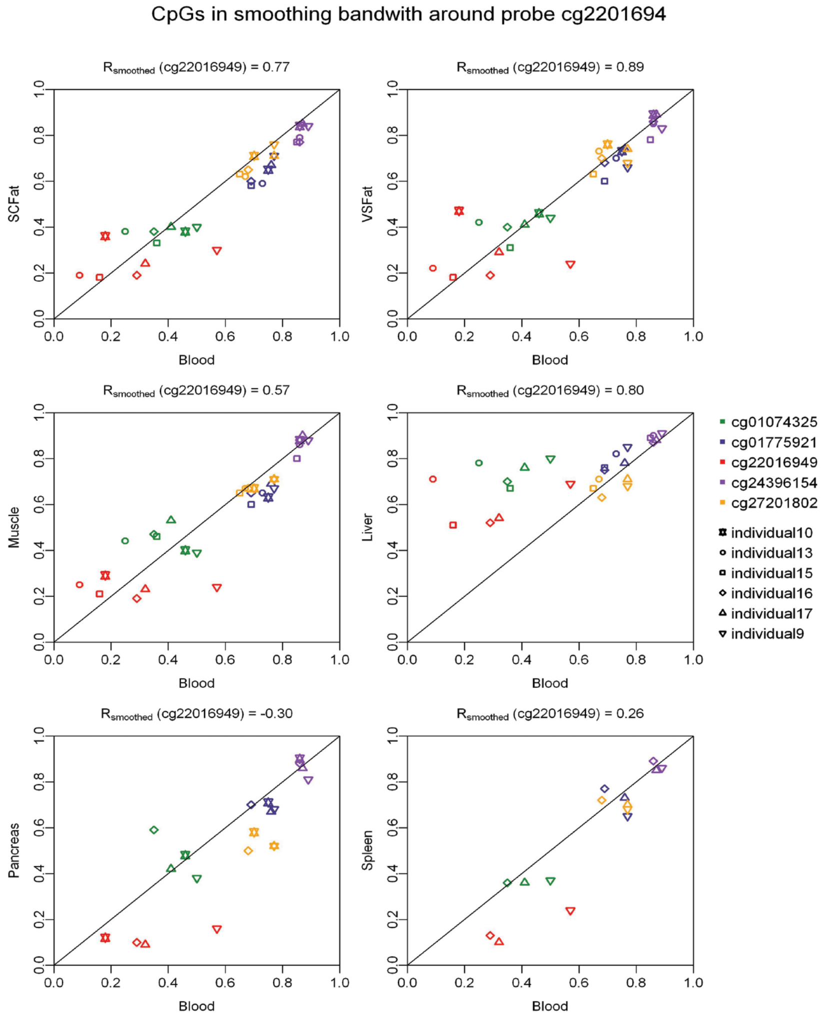


Figure 6.4: An example of smoothing tissue correlations over all CpG sites within a region of 2000 bp. The scatter plots show DNA methylation (in β-value) of blood (x-axes) against each other tissue (y-axes) per individual (plotted symbol) per CpG site (color) in the smoothing bandwidth around probe cg2201694. The smoothed tissue correlation of cg2201694 is given above each plot. The x = y line is given in black for reference.

Table 6.3A: Frequencies of CpG sites with strong and weak correlation between blood and the tissues in CpG rich and CpG poor areas

CpG density[a]	Correlated[b]	Amount (percentage) of CpG sites correlated between blood and each tissue[c]						
		SCFat	VSFat	Muscle	Liver	Spleen	Pancreas	All tissues
CpG poor	Weak	148863 (93.81 %)	150683 (94.96 %)	153464 (96.71 %)	147441 (92.92 %)	134263 (84.61 %)	146658 (92.42 %)	109983 (69.31 %)
CpG poor	Strong	9818 (6.19 %)	7998 (5.04 %)	5217 (3.29 %)	11240 (7.08 %)	24418 (15.39 %)	12023 (7.58 %)	48698 (30.69 %)
CpG rich	Weak	206057 (93.85 %)	208456 (94.94 %)	212572 (96.82 %)	204114 (92.97 %)	185553 (84.51 %)	203143 (92.52 %)	152410 (69.42 %)
CpG rich	Strong	13501 (6.15 %)	11102 (5.06 %)	6986 (3.18 %)	15444 (7.03 %)	34005 (15.49 %)	16415 (7.48 %)	67148 (30.58 %)

Table 6. 3B: Frequencies of CpG sites from CpG rich areas with a strong correlation between blood and the tissues

	Amount (percentage) of CpG sites correlated between blood and each tissue[d]						
Correlated[b]	SCFat	VSFat	Muscle	Liver	Spleen	Pancreas	All tissues
Strong with 1 tissue[e]	5447 (1.44 %)	3606 (0.95 %)	1986 (0.53 %)	6714 (1.78 %)	20266 (5.36 %)	8255 (2.18 %)	46274 (12.23 %)
Strong with 2 tissues[e]	4224 (1.12 %)	3781 (1.00 %)	2257 (0.60 %)	5117 (1.35 %)	9041 (2.39 %)	5144 (1.36 %)	14782 (3.91 %)
Strong with 3 tissues[e]	2092 (0.55 %)	2007 (0.53 %)	1228 (0.32 %)	2010 (0.53 %)	2985 (0.79 %)	1738 (0.46 %)	4020 (1.06 %)
Strong with 4 tissues[e]	875 (0.23 %)	847 (0.22 %)	679 (0.18 %)	776 (0.21 %)	945 (0.25 %)	586 (0.15 %)	1177 (0.31 %)
Strong with 5 tissues[e]	491 (0.13 %)	489 (0.13 %)	464 (0.12 %)	455 (0.12 %)	396 (0.10 %)	320 (0.08 %)	523 (0.14 %)
Strong with 6 tissues[e]	372 (0.10 %)	372 (0.10 %)	372 (0.10 %)	372 (0.10 %)	372 (0.10 %)	372 (0.10 %)	372 (0.10 %)
Strong mesodermal[f]	1355 (0.36 %)	1262 (0.33 %)	888 (0.23 %)				1682 (0.44 %)
Strong endodermal[f]				3994 (1.06 %)	5405 (1.43 %)	3956 (1.05 %)	6472 (1.71 %)

a: CpG rich area: 3 or more additional probes in smoothing bandwidth (2000 bp)
b: Weak correlation: r < 0.75; Strong correlation: r ≥ 0.75;
c: Percentages are of all CpG sites within CpG poor areas (158,681 CpG sites) or within CpG rich areas (219,558 CpG sites)
d: Percentages are of all CpG sites examined (378,239 CpG sites)
e: CpG sites in CpG rich area with strong correlation between blood and number of tissues
f: CpG sites in CpG rich area with strong correlation between blood and more than one mesodermal / endodermal tissue

blood and the endodermal tissues compared with the mesodermal tissues (Table 6.3B). Although, a strong tissue correlation was observed between blood and one tissue exclusively for most CpG sites, more than 6,000 CpG sites showed a strong tissue correlation with three or more tissues, and 372 CpG sites with all tissues (Table 6.3B and, for each combination of tissues, Supplementary table S6.1). We then counted the amount of these tissue correlations per genomic fragment of 100,000 bp long (100K block) and scanned for genomic areas enriched for tissue correlation in a genome wide plot (Figure 6.5). We found that genomic areas with a higher amount of correlated probes (67,148 probes distributed over 11,080 100K blocks, plotted in blue) corresponded to areas with more measurements in general (219,558 probes distributed over 12,060 100K blocks, plotted in black). The proportion of correlated probes per 100K block (≥ 15 measurements; 162,014 probes distributed over 4,907 100K blocks, plotted in green) revealed no extended area with specific enrichment of tissue correlation, although there were 100K blocks with low (< 0.10) and with a high (> 0.50) proportion of correlated probes. In general one third of probes displayed a tissue correlation irrespective of measurement density, with a similar genome wide distribution of the amount of probes at which methylation of blood was correlated with one or two internal tissues (61,052 probes distributed over 11,077 100K blocks, plotted in purple), or with multiple tissues (5,674 probes distributed over 3,850 100K blocks, plotted in orange). At last, the distribution of the 372 probes correlated between blood and all internal tissues (distributed over 288 100K blocks, plotted in red) seems less related to measurement density.

Characteristics of correlated CpG sites

We inspected the distributions of CpG sites across the categories of mean methylation, methylation variation, and tissue differences to investigate whether strong tissue correlations could be marked by such characteristics of DNA methylation. We focused this inspection on CpG sites from CpG rich areas (219,558 sites), as their correlations represent a weighted average over multiple CpG sites. We observed no difference in these distributions between CpG sites with a strong tissue correlation compared with all CpG sites from CpG rich areas in any of the tissues (Tables 6.4 and 6.5).

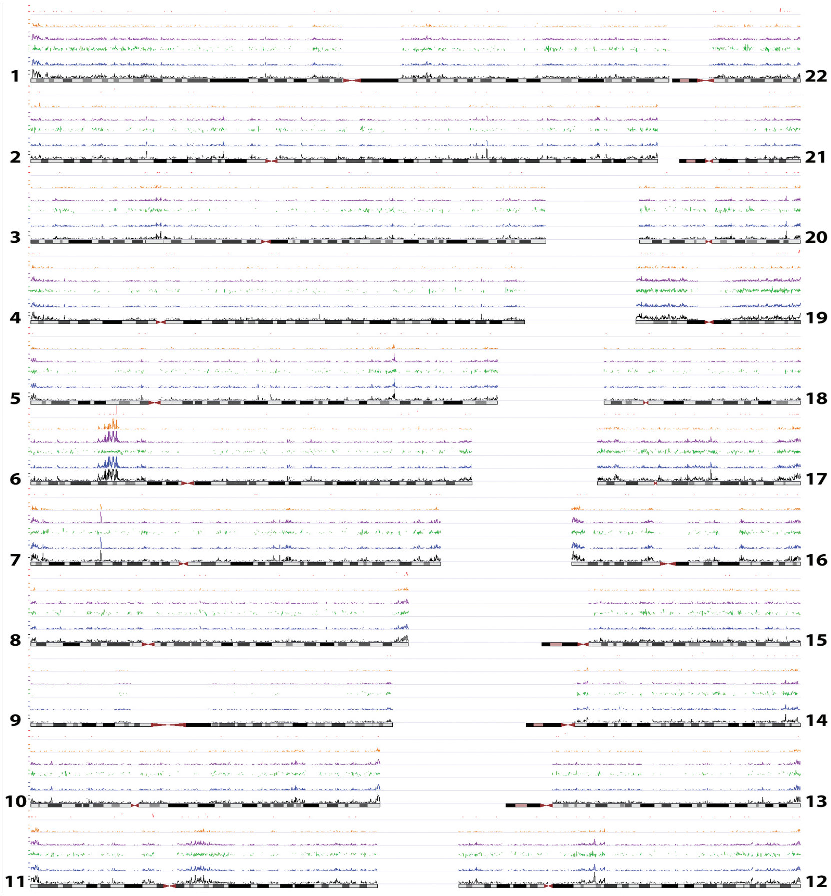


Figure 6.5: Plot of correlation of DNA methylation between blood and internal tissues (subcutaneous fat, visceral fat, skeletal muscle, liver, pancreas, and spleen) across the human genome (assembly: hg 18 / NCBI 36). Lines above each chromosome represent frequency information per unit of 100,000 bp long (100K block) of the number of probes with 3+ additional CpGs in the 2,000 bp smoothing bandwidth (black line), the number of probes with strong tissue correlation (blue line), the proportion of tissue correlated probes for blocks containing ≥ 15 probes (green line), the number of probes with tissue correlation between blood and 1 or 2 tissues (purple line), between blood and 3 to 5 tissues (orange line) and between blood and all 6 tissues (red line). The plot was made using the genome graph function of the UCSC genome browser [150] (website: http://genome.ucsc.edu/).

Table 6.4: Frequency distributions of mean DNA methylation and methylation variation for CpGs with strong and weak correlations

	CpG sites		Distribution (%) over mean methylation categories[a]					Distribution (%) over variation categories[b]		
Tissue	Category	n	0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8–1	0 – 0.05	0.05 – 0.15	0.15 – 0.9
Blood	Strong CpG rich[c]	67148	53.9	8.9	5.6	10.9	20.6	44.5	46.4	9.2
	All CpG rich[d]	219558	56.0	8.9	5.5	10.2	19.4	46.0	44.9	9.1
	All CpG poor[e]	158681	6.9	5.0	7.7	25.8	54.6	19.6	66.9	13.5
SC Fat	Strong CpG rich[c]	13501	55.6	8.5	8.9	11.8	15.2	49.8	46.7	3.5
	All CpG rich[d]	219558	56.2	8.8	8.2	10.8	15.9	50.8	46.0	3.2
	All CpG poor[e]	158681	5.9	5.7	12.5	28.2	47.8	33.9	62.4	3.7
VS Fat	Strong CpG rich[c]	11102	52.6	9.2	7.9	12.7	17.7	46.4	39.4	14.2
	All CpG rich[d]	219558	54.3	9.4	7.6	11.0	17.6	50.1	36.9	13.1
	All CpG poor[e]	158681	5.1	4.6	10.2	28.6	51.5	28.0	50.7	21.4
Muscle	Strong CpG rich[c]	6986	56.3	9.2	8.6	10.0	15.8	68.0	30.1	1.9
	All CpG rich[d]	219558	56.2	9.0	8.0	10.3	16.5	68.3	29.9	1.8
	All CpG poor[e]	158681	6.4	7.7	12.7	24.7	48.6	46.7	50.4	2.9
Liver	Strong CpG rich[c]	15444	51.1	10.9	8.6	10.9	18.5	63.3	33.9	2.9
	All CpG rich[d]	219558	52.1	10.7	8.1	10.5	18.6	64.6	32.8	2.6
	All CpG poor[e]	158681	4.8	5.1	10.5	26.1	53.5	48.2	47.6	4.2
Spleen	Strong CpG rich[c]	34005	54.2	8.7	6.8	11.7	18.6	65.5	29.6	4.9
	All CpG rich[d]	219558	55.4	8.7	6.7	10.7	18.5	66.3	29.1	4.7
	All CpG poor[e]	158681	5.6	4.6	9.0	27.6	53.1	51.2	42.0	6.8
Pancreas	Strong CpG rich[c]	16415	55.8	9.0	8.0	10.3	16.9	61.3	32.5	6.1
	All CpG rich[d]	219558	55.8	8.7	7.5	10.3	17.7	62.2	31.8	5.9
	All CpG poor[e]	158681	5.8	6.1	11.4	29.6	47.0	33.4	56.2	10.4

a: Lower to upper boundary of mean methylation for categories expressed in β-value

b: Lower to upper boundary of size of SD interval for categories expressed in β-value

c: CpG sites in CpG rich areas with a strong correlation ($r > 0.75$) between blood and at least one tissue, for blood this means all such CpG sites

d: All CpG sites in CpG rich areas, which had 3 or more additional CpG sites available for smoothing within the 2000 bp bandwidth

e: All CpG sites in CpG poor areas, which had 2 or less additional CpG sites available for smoothing within the 2000 bp bandwidth

Table 6.5: Frequency distributions of tissue differences in average DNA methylation for correlated and uncorrelated CpG sites

	CpG sites		Distribution (%) over methylation difference categories[a]			
Blood vs Tissue	Category	n	Insignificant	0 – 0.05	0.05 – 0.20	0.20 – 1
SC Fat	Strong CpG rich[b]	13501	72.5	16.2	9.1	2.2
	All CpG rich[c]	219558	73.0	18.0	7.5	1.5
	All CpG poor[d]	158681	59.9	26.9	10.9	2.2
VS Fat	Strong CpG rich[b]	11102	83.5	13.5	2.8	0.1
	All CpG rich[c]	219558	81.2	14.6	3.8	0.4
	All CpG poor[d]	158681	73.9	20.2	5.2	0.6
Muscle	Strong CpG rich[b]	6986	58.5	25.2	11.6	4.7
	All CpG rich[c]	219558	59.3	23.4	12.6	4.6
	All CpG poor[d]	158681	46.5	30.9	16.7	5.8
Liver	Strong CpG rich[b]	15444	70.5	14.6	10.9	4.0
	All CpG rich[c]	219558	70.0	19.6	8.3	2.2
	All CpG poor[d]	158681	59.2	26.5	11.4	2.8
Spleen	Strong CpG rich[b]	34005	51.9	36.6	10.7	0.9
	All CpG rich[c]	219558	53.1	37.8	8.5	0.6
	All CpG poor[d]	158681	51.0	39.8	8.6	0.6
Pancreas	Strong CpG rich[b]	16415	92.3	3.4	3.2	1.0
	All CpG rich[c]	219558	91.6	4.6	3.0	0.8
	All CpG poor[d]	158681	90.1	5.6	3.4	0.9

a: Lower to upper boundary of methylation difference for categories expressed in β-value, when $p_{T\text{-test}} < 0.05$

b: CpG sites in CpG rich areas with a strong correlation (r > 0.75) between blood and at least one tissue, for blood this means all such CpG sites

c: All CpG sites in CpG rich areas, which had 3 or more additional CpG sites available for smoothing within the 2000 bp bandwidth

d: All CpG sites in CpG poor areas, which had 2 or less additional CpG sites available for smoothing within the 2000 bp bandwidth

Different distributions were observed between CpG sites from CpG rich areas and CpG poor areas. Mean methylation in CpG rich areas displayed a similar bimodal distribution as for all 378,239 CpG sites (Supplementary figure S6.5), although the peak at low methylation (ß < 0.2) was substantially higher (> 50 % of CpG sites) than the peak at high methylation (ß ≥ 0.8; 15 % – 20 % of CpG sites). In contrast, mean methylation in CpG poor areas displayed a skewed shape, with the tail at low methylation (< 7 % of CpG sites) and the peak at high methylation (> 50 % of CpG sites). These distributions were of similar shape in al tissues (Table 6.4). Methylation variation in all tissues was substantially higher in CpG poor areas compared with CpG rich areas (Table 6.4). Tissue differences also appeared more pronounced in CpG poor areas compared with CpG rich areas, except for spleen and pancreas (Table 6.5).

Finally, we investigated the possibility of identifying CpG sites with strong tissue correlations by the genomic function, structure, or genetic sequence of their surrounding area. We first inspected frequency distributions of functional areas and observed differences in the function of the location of CpG rich and CpG poor areas, but not of CpG sites with strong or weak tissue correlations, either in CpG rich or in CpG poor areas (Supplementary table S6.2). Comparing CpG sites with the strongest and the weakest tissue correlations (6539 and 5837 sites, respectively), we tested and found that neither repetitive DNA, nor RefSeq defined genetic components, nor CpG islands, nor most regulatory features were associated with strong tissue correlations (Table 6.6). We further tested and found that CpG sites with strong tissue correlations do not have different DNA structure or sequence compared with CpG sites with weak tissue correlations. Most differences between strongly and weakly correlated CpG sites were observed for epigenomic features (Table 6.6). Details of features with nominally significant ($p < 0.05$) differences between CpG sites with strong and weak tissue correlations are given in Supplementary table S6.3.

Discussion

In this study we surveyed DNA methylation at 380,000 CpG sites distributed throughout the genome across seven tissues of six individuals, with the aim to uncover loci at which DNA methylation in blood may mark that of inaccessible tissues [202]. Resembling results of previous studies [100–102], variation in genome wide DNA methylation patterns was largely determined by tissue differences. However, in line with results from recent studies [195,234], a substantial subset of individual CpG sites, showed similar

Table 6.6: Exploration of associations between tissue correlation and structural genomic and epigenomic features, or genomic location

Annotations by (test type)[a]	Genomic or epigenomic features investigated for association with tissue correlation[b]		
	Feature name	Investigated	Associated
epiGRAPH (wilcoxon)	Chromosome Organisation	27	3
	DNA Structure	63	0
	Epigenome and Chromatin Structure	107	23
	Conserved TFBS	260	8
	Repetitive DNA	162	0
Illumina (chi square)	Regulatory feature	8	2
	Relation to CpG island	5	0
	RefSeq group	6	0

a: Features and relative genomic locations using the epigraph web tool, or the annotations of the platform manufacturer

b: The number of features investigated vs the number of features with a nominally significant ($p < 0.05$) difference between CpG sites with strong and weak tissue correlations

average DNA methylation or a high correlation coefficient between blood and an internal tissue. High correlation coefficients were neither associated with levels of average DNA methylation in both tissues nor with methylation variation, nor with methylation differences between the tissues. Genetic sequence, structure and genomic location also appeared unassociated with tissue correlations. Hence, although we found loci for which blood can represent epigenetic marks of inaccessible tissue, we did not find a distinguishing feature by which such loci can be recognized. Still, cataloging the loci with methylation similarities between tissues, both those with comparable methylation levels and those with a high correlation of methylation variation, will indicate the loci at which an accessible tissue like blood can be used as a marker tissue for disease related tissues that are inaccessible. Using a marker tissue for inaccessible tissues may help the design of epigenetic epidemiological studies and relate results to common diseases beyond the context of the tissue measured [98,99,202].

Despite the striking differences in genome-wide methylation patterns between the tissues, we also observed that average DNA methylation at individual CpG sites was more often similar than different between blood and each tissue, which was also reported on in a previous study [100]. Further, roughly a quarter of the CpG sites investigated showed a similar average DNA methylation level across multiple tissues (frequently hypo- or hyper-methylated). In line with previous studies [100–102], these results indicate the existence of many loci at which methylation levels will be similar between blood and an inaccessible tissue of interest. As tissue differences in methylation levels at CpG sites are often interpreted as indicative of a difference in epigenetic regulation of gene expression [101] and / or cell type differentiation [88,102]. Conversely, CpG sites with a similar methylation level between tissues might indicate a similar regulatory role of the epigenetic mechanism at the locus for both tissues. However, the aim of this study was to uncover loci at which the methylation status of blood marks that of inaccessible tissues for which the correlation of methylation variation between the tissues will be more relevant than the actual methylation level in either tissue.

We observed that a substantial subset of CpG sites showed a strong correlation between blood and another tissue. Previous studies have reported on inter tissue correlations at specific candidate loci [133,134,195], and here we present more of such data on a genome wide scale [234]. The majority of strong correlations were found between blood and one tissue exclusively, but a still sizeable minority showed a strong correlation across multiple tissues. A few CpG sites were strongly correlated across all tissues, however the most logical explanation for such seemingly soma-wide correlation of DNA methylation will be unknown genetic background variation, unless there is a biological reason to assume otherwise, such as genomic imprinting. Thus, for epigenetic case-control studies, not the similarity of average methylation, but a high correlation of methylation variation between the tissues will define loci at which blood can be a marker tissue for inaccessible tissues that are more relevant for the disease [119,195,234].

We focused our survey on correlated CpG sites that were smoothed over a minimum of four CpG sites, which were inevitably located in CpG rich areas. such CpG sites with a strong tissue correlation after smoothing are more likely to produce reliable assays in epigenetic epidemiology potentially marking the epigenetics of an internal tissue. However, the percentage of CpG sites with a strong correlation was similar between CpG rich and CpG poor areas. Thus, there is no reason to assume that by definition CpG poor areas will provide less suitable targets. This could be ascertained by validating the strong tissue correlations of CpG poor areas in a study with bigger sample sizes. Since DNA methylation changes progressively during cell differentiation [30] the differences between cell lineages that separated early during development are likely more pronounced. We therefore anticipated more tissue correlations between blood and the other mesodermal tissues compared with blood and the endodermal tissues [195]. However, we observed the contrary, which suggests that fewer DNA methylation changes occur before the separation of the germ layers than occur between the different cell lineages within a germ layer. On the other hand, there were fewer individuals sampled for each of the endodermal tissues, the influence of which cannot be estimated. Tissue correlations, either between blood and one tissue or across multiple tissues, were unrelated to mean methylation levels, methylation variation, or tissue differences at the CpG site, indicating that such characteristics of DNA methylation itself cannot be used to identify likely candidate CpG sites with strong correlation between marker tissue and disease related tissue. Correlation between DNA methylation in blood with that in other tissues appears to be complex and dependent on both locus and tissue, as previously suggested [195].

Design of epigenetic studies and interpretation of their results, would be greatly enhanced if genetic or genomic characteristics could be revealed by which the potential for tissue correlation of a locus can be recognized [119,195]. A good source for such characteristics may be related to the function of DNA methylation [24]. Repetitive DNA tends to be hyper-methylated in all tissues [39], which implies that DNA methylation at repetitive elements may be more discriminative between individuals [241]. In contrast, DNA methylation at CpG island shores was recently demonstrated as especially discriminative between tissues [129]. It may be reasonable to expect CpG sites with a strong correlation between tissues to be located more frequently in repetitive elements and less frequently in CpG island shores. However, in neither of these features did we observe a different frequency of appearance for the most correlated CpG sites compared with the least correlated CpG sites. Similarly, we also did not observe any difference between CpG sites with strongest and weakest correlations with respect to their genetic sequence motifs, their genome structure or their location relative to genomic functional elements, for all of which the epigenetic state has been repeatedly related to cellular gene expression potential [21,89]. These results would suggest that correlation of DNA methylation between tissues may not be related to its functional aspects. On the other hand, our exploratory results indicate that a subset of epigenomic features in blood, such as histon modifications, particularly methylation at various lysine residues of Histon H3, and associations with RNA polymerase II, may be related to tissue correlations. This apparent contradiction cannot be resolved in our current study. To elucidate what causes a strong correlation of DNA methylation between tissues at certain loci, and by which features such loci can be identified will require substantial research efforts.

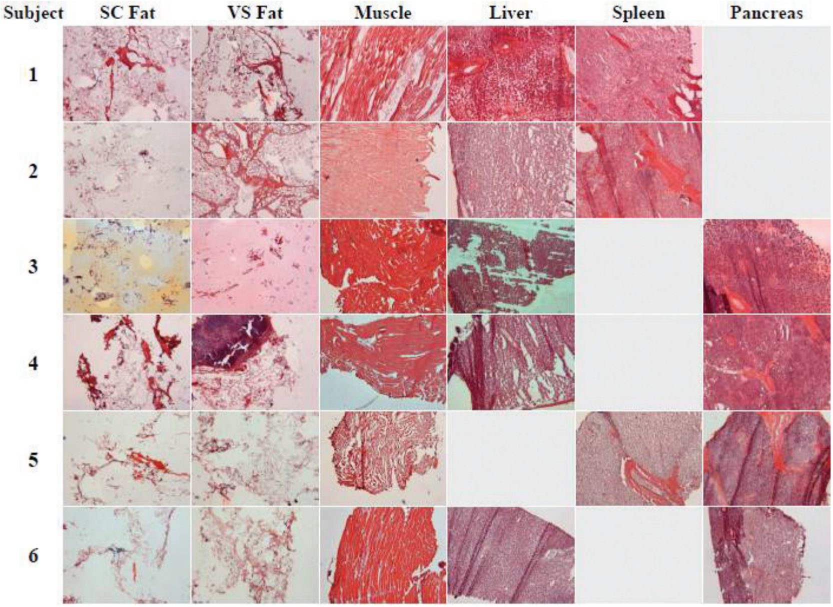
We compared CpG methylation of DNA from tissue samples collected during autopsy. DNA methylation is a covalent chemical bond that is stable over many years in stored DNA [195,197], and was shown to change little within the first few days after death [238], hence we set the maximum post mortem interval to 24 hours after death. In blood we found no indication that DNA methylation patterns were dramatically altered within these 24 hours after death, and there is no reason to assume this would be different for the other tissues. Thus the extrapolation of our observations to living individuals seems reasonable. Between chip variation of measurements was shown to be small and individuals whose DNA was analyzed on the same chip did not form clusters in the PCA. Data from probes with a known SNP in their 50 bp target sequence [238] was excluded from our analyses. Therefore technical artifacts seem an unlikely explanation for most of the observed tissue correlations, although the influence of (unknown) genetic variation in cis cannot be excluded [242,243]. To accommodate the limitations of the small sample size of our study we employed a smoothing algorithm to create a median of the observed effects across the CpG sites within an area the size of 2000 bases. This will make the observation indicative of more robust locally carried effects, a method with demonstrated validity [135,233,244]. This will only work well at CpG sites if the recorded value can be internally validated by effects at sufficient nearby CpG sites. Thus, our observations at CpG sites in CpG poor areas lack the confidence granted by a sufficient sample size, and unfortunately have no other method available for increasing the robustness of the observed effect. This compelled us to adjust the comparison of strong and weak correlations to the CpG richer areas, which have different methylation characteristics (mean, variation and tissue differences), as observed in this study and in previous literature [90].

In this study we report that DNA methylation in blood may mark that of an internal tissues at a substantial subset of loci, which can be exploited for choosing assays and for interpreting results of epigenetic research. This capacity was established by strong correlation coefficients, which were complex in nature, depending on both locus and tissues. Strong correlations were unrelated to characteristics of DNA methylation, nor to the genetic sequence and genomic features surrounding the CpG site. Blood is the only tissue collected in most current human biobanks, due to its simple and cheap sampling techniques with low invasiveness. Other tissues that can also be collected in sufficient quantities for large populations with simple, inexpensive, and low invasive techniques are buccal swabs [245], epidermis [108], bladder lining from urine, and colonic mucosa from stool [195]. Such tissues may be used as marker tissues for the internal disease related tissues at different loci than are marked with blood. Although the effort to assess the usability maker tissues is substantial, epigenetic research on the common diseases will benefit greatly from the capacity to interrogate as many loci as possible, providing meaningful results through use of the right marker tissue (or more than one tissue for extra robustness) for that locus in combination with the disease related tissue.

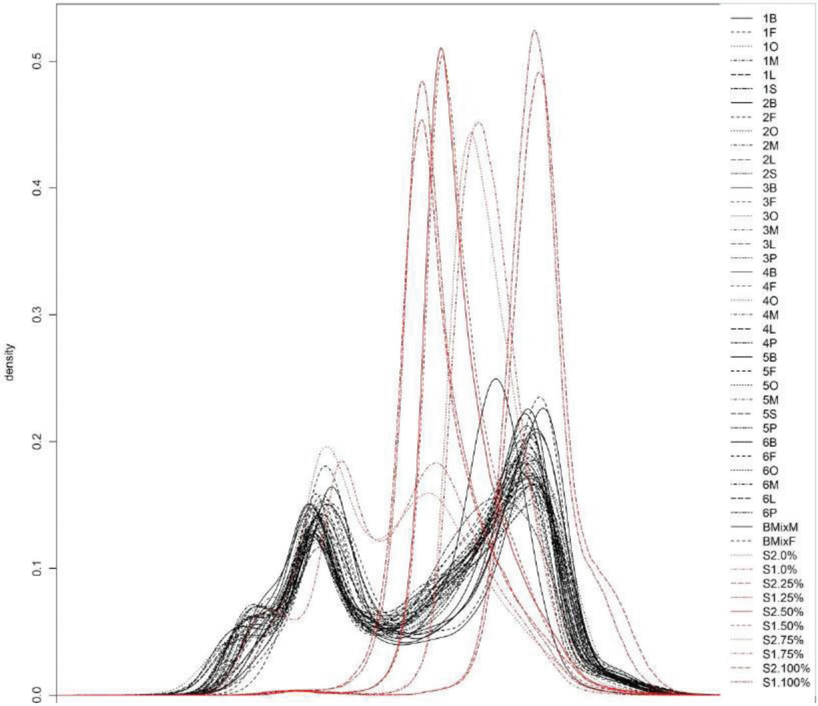
Acknowledgement

We would like to thank R. Guijt, T. Alberts, and E. Dreef for helping us obtain samples during the post mortem investigations, and W. de Dijcker for his help in obtaining and processing the samples. The research leading to these results has received funding from the Netherlands Heart Foundation (2006B083), the European Union’s Seventh Framework Program IDEAL (FP7/2007-2011) under grant agreement n° 259679, the Netherlands Consortium for Healthy Ageing (Grant 05060810) in the framework of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research, and Grant Number R01AG042190 from the National Institute On Aging.

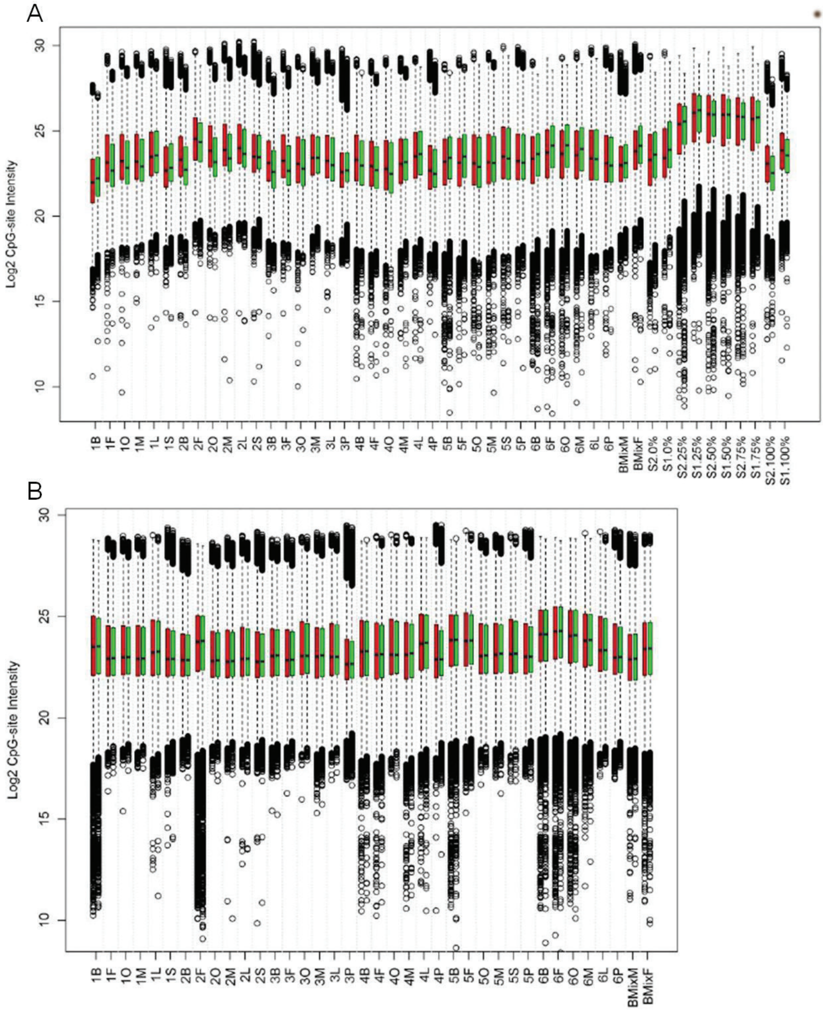
Supplementary Figures



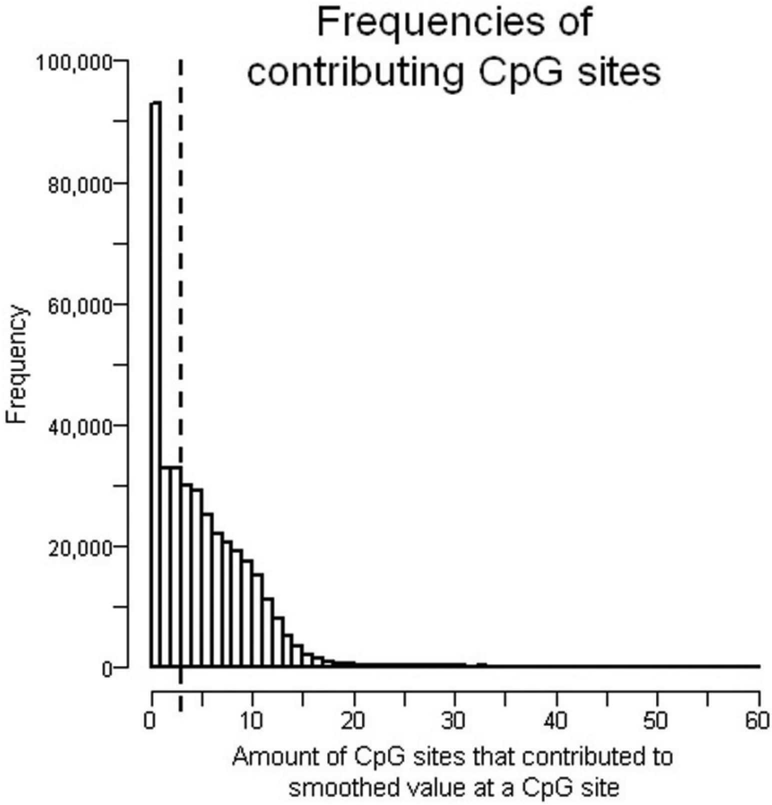
Supplementary Figure S6.1: Microscopy images of tissue coupes. Coupes are stained with HE, staining proteins and cytoplasm red and nuclei blue. SC Fat: subcutaneous fat; VS Fat: visceral fat. Except for visceral fat, the heterogeneity of the tissues appears similar between the individual



Supplementary Figure S6.2: Density plots of the raw M-values. Different methylation patterns are observed across all CpG sites between the tissue samples (black lines) and the methylation titer series (red lines). B:Blood, F:Fat, O:Visceral fat, M:Skeletal muscle, L:Liver, S:Spleen, P:Pancreas. 1-6: invididual 1 to 6, BMixM/F: Blood mixture Male/Female, S1/2.x%: Series1 or 2, x% methylated.



Supplemental Figure S6.3: Boxplots showing the intensity distribution (y-axis) of the two colour channels of the samples (x-axis) before (A) and after (B) normalization.
B:Blood, F:Fat, O:Visceral fat, M:Skeletal muscle, L:Liver, S:Spleen, P:Pancreas.
1-6: invididual 1 to 6, BMixM/F: Blood mixture Male/Female, S1/2.x%: Series1 or 2, x% methylated.



Supplementary figure S6.4: Distribution plot of the amount of CpG sites over which the correlation at a CpG site was averaged during smoothing. The vertical line represents the boundary between what are called in this study the CpG sites in CpG rich areas (3 or more additional CpG sites, thus correlation being a median over 4 CpGs) and the CpG sites in CpG poor areas.

Supplementary tables

Table S6.1: Frequency distribution of CpG sites in CpG rich areas with strong correlation (r ≥ 0.75) between blood and each combination of tissues

		Combination of tissues					
Amount of tissues	**Amount of CpG sites**	**SCFat**	**VSFat**	**Muscle**	**Liver**	**Spleen**	**Pancreas**
0	311091	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx	xxxxxx
1	5447	x					
1	3606		x				
1	1986			x			
1	6714				x		
1	20266					x	
1	8255						x
2	794	x	x				
2	420	x		x			
2	327		x	x			
2	853	x			x		
2	378		x		x		
2	303			x	x		
2	1505	x				x	
2	1648		x			x	
2	894			x		x	
2	2516				x	x	
2	652	x					x
2	634		x				x
2	313			x			x
2	1067				x		x
2	2478					x	x
3	141	x	x	x			
3	178	x	x		x		
3	112	x		x	x		
3	52		x	x	x		
3	461	x	x			x	
3	206	x		x		x	
3	182		x	x		x	
3	430	x			X	X	
3	343		x		X	X	

		Combination of tissues					
Amount of tissues	Amount of CpG sites	SCFat	VSFat	Muscle	Liver	Spleen	Pancreas
3	177			X	X	X	
3	93	X	X				x
3	69	x		x			x
3	83		X	X			x
3	161	x			x		x
3	100		x		x		x
3	46			X	X		x
3	241	x				X	X
3	374		x			X	X
3	160			x		X	X
3	411				X	X	X
4	74	X	X	X	X		
4	139	X	X	X		x	
4	195	X	X		X	X	
4	110	x		X	X	X	
4	73		X	X	X	X	
4	41	X	X	X			x
4	40	X	X		x		x
4	51	x		x	X		x
4	26		X	X	X		x
4	103	X	X			X	X
4	38	x		x		X	X
4	80		X	X		X	X
4	84	x			x	x	X
4	76		x		X	X	X
4	47			X	X	X	X
5	203	x	x	X	X	X	
5	127	X	X	X	X		x
5	68	X	X	X		X	X
5	59	X	X		X	X	X
5	34	x		X	X	X	X
5	32		x	X	X	X	X
6	372	X	X	X	X	X	X

The green background designates the mesodermal tissues and the blue background designates the endodermal tissues

Supplementary table S6.2: Distributions of genomic locations for CpG sites in CpG poor and rich areas with weak and strong tissue correlations

Genomic location[a]	Sites from CpG poor areas		Sites from CpG rich areas		
	Weak in 6T[b] (n = 109983)	Strong in 1T – 6T[b] (n = 48698)	Weak in 6T[b] (n = 152410)	Strong in 1T – 2T[b] (n = 61056)	Strong in 3T – 6T[b] (n = 6092)
In relation to RefSeq gene					
none	41173 (37.4 %)	18257 (37.5 %)	21703 (14.2 %)	8834 (14.5 %)	900 (14.8 %)
1stExon	1154 (1.0 %)	490 (1.0 %)	11332 (7.4 %)	4665 (7.6 %)	423 (6.9 %)
3 UTR	6393 (5.8 %)	2944 (6.0 %)	2716 (1.8 %)	1057 (1.7 %)	93 (1.5 %)
5 UTR	7694 (7.0 %)	3331 (6.8 %)	15767 (10.3 %)	6367 (10.4 %)	604 (9.9 %)
Body	46985 (42.7 %)	20886 (42.9 %)	40377 (26.5 %)	16096 (26.4 %)	1639 (26.9 %)
TSS 1500	5453 (5.0 %)	2368 (4.9 %)	32681 (21.4 %)	13112 (21.5 %)	1360 (22.3 %)
TSS 200	1131 (1.0 %)	422 (0.9 %)	27834 (18.3 %)	10925 (17.9 %)	1073 (17.6 %)
In relation to regulatory feature and cell type specificity (CTS)					
none	93036 (84.6 %)	41855 (85.9 %)	72808 (47.8 %)	30403 (49.8 %)	3089 (50.7 %)
Gene Associated	389 (0.4 %)	161 (0.3 %)	318 (0.2 %)	144 (0.2 %)	12 (0.2 %)
Gene Associated (CTS)	720 (0.7 %)	276 (0.6 %)	376 (0.2 %)	166 (0.3 %)	15 (0.2 %)
Non Gene Associated	56 (0.1 %)	17 (0.0 %)	710 (0.5 %)	263 (0.4 %)	31 (0.5 %)
Non Gene Associated (CTS)	37 (0.0 %)	16 (0.0 %)	75 (0.0 %)	36 (0.1 %)	3 (0.0 %)
Promoter Associated	2622 (2.4 %)	998 (2.0 %)	50386 (33.1 %)	18854 (30.9 %)	1802 (29.6 %)
Promoter Associated (CTS)	816 (0.7 %)	276 (0.6 %)	2458 (1.6 %)	903 (1.5 %)	85 (1.4 %)
Unclassified	4539 (4.1 %)	1892 (3.9 %)	12712 (8.3 %)	5033 (8.2 %)	512 (8.4 %)
Unclassified (CTS)	7768 (7.1 %)	3207 (6.6 %)	12567 (8.2 %)	5254 (8.6 %)	543 (8.9 %)
In relation to CpG island					
Open sea	68053 (61.9 %)	30002 (61.6 %)	23978 (15.7 %)	9655 (15.8 %)	936 (15.4 %)
N Shelf	10553 (9.6 %)	4560 (9.4 %)	2638 (1.7 %)	1125 (1.8 %)	104 (1.7 %)
N Shore	8783 (8.0 %)	3937 (8.1 %)	25944 (17.0 %)	10286 (16.8 %)	1101 (18.1 %)
Island	6388 (5.8 %)	2967 (6.1 %)	77380 (50.8 %)	30901 (50.6 %)	3016 (49.5 %)
S Shelf	9724 (8.8 %)	4330 (8.9 %)	1979 (1.3 %)	846 (1.4 %)	83 (1.4 %)
S Shore	6482 (5.9 %)	2902 (6.0 %)	20491 (13.4 %)	8243 (13.5 %)	852 (14.0 %)

a: Information as supplied by manufacturer (Illumina, San Diego, USA)

b: 6T: in all six tissues; 1T – 6T: in one or more tissues; 1T – 2T: in one or two tissues; 3T – 6T: in 3 or more tissues

Supplementary table S6.3: Differences in structural features between CpG sites with the strongest and the weakest tissue correlations

Attribute Group Name	Attribute Name	P^{a}_{wilcoxon}	Weakest (n = 5837)		Strongest (n = 6539)	
			Mean	SD	Mean	SD
Chromosome Organisation	gieStain gpos50 overlap Average Size	.010	3911567	1769849	4131498	1760516
	gieStain gpos75 overlap Regions Count	.008	3.73	13.15	3.14	12.12
	gieStain gpos75 overlap Total Length	.008	74.70	262.92	62.70	242.44
Epigenome and Chromatin Structure	Overlap Regions Count	.003	54.19	77.77	58.57	80.98
	Overlap Total Length	.004	32.78	44.43	35.20	45.93
	tissue H1 overlap Regions Count	.003	27.15	41.54	29.55	43.39
	tissue H1 overlap Total Length	.003	27.15	41.54	29.55	43.39
	tissue imr90 overlap Regions Count	.018	27.03	41.81	29.03	43.50
	tissue imr90 overlap Total Length	.018	27.03	41.81	29.03	43.50
	cTissueHes hues8 oMeth ratio	.037	0.20	0.33	0.21	0.33
	chromMod CTCF overlap Regions Count	.007	22.71	104.89	20.88	112.90

Attribute Group Name	Attribute Name	$P^{a}_{wilcoxon}$	Weakest (n = 5837)		Strongest (n = 6539)	
			Mean	SD	Mean	SD
	chromMod CTCF overlap Total Length	.006	107.92	276.97	94.20	259.81
	chromMod H2A Z overlap Regions Count	.002	72.97	171.05	65.12	156.06
	chromMod H2A Z overlap Total Length	.003	298.63	417.45	277.08	408.61
	chromMod H3K27me1 overlap Average Size	.031	24.33	0.45	24.28	0.45
	chromMod H3K36me1 overlap Regions Count	.024	10.41	24.44	9.28	22.64
	chromMod H3K36me1 overlap Total Length	.027	107.28	266.07	97.13	253.34
	chromMod H3K4me3 overlap Regions Count	$5.4*10^{-05}$	379.11	775.23	336.92	714.18
	chromMod H3K4me3 overlap Total Length	$9.9*10^{-05}$	547.73	462.54	516.12	463.05
	chromMod H3K79me1 overlap Regions Count	.003	13.88	27.62	12.48	26.52
	chromMod H3K79me1 overlap Total Length	.005	138.20	292.36	126.51	282.59
	chromMod PolII overlap Regions Count	$1.1*10^{-06}$	51.54	140.81	44.68	136.63
	chromMod PolII overlap Total Length	$8.1*10^{-07}$	252.51	394.28	219.14	375.02

a: Unadjusted p-value from Wilcoxon (epigraph webtool) or Chi Square test, only attributes with nominally significant ($p < 0.05$) differences are given

Supplementary table S3 (continued)

Attribute Group Name	Attribute Name	P^a_{wilcoxon}	Weakest (n = 5837) Mean	SD	Strongest (n = 6539) Mean	SD
Epigenome and Chromatin Structure	Overlap Average Size	.006	0.48	0.01	0.48	0.01
	Overlap Regions Count	$9.5*10^{-06}$	823.96	1041.14	761.10	981.93
	Overlap Total Length	.017	922.32	216.52	912.44	233.22
Conserved TFBS	cNameV ahrarnt 02 oZscore	.017	1.98	0.28	2.33	0.49
	cNameV creb 02 oZscore	.004	1.76	0.13	2.36	0.42
	cNameV elk1 01 oZscore	.030	2.58	0.59	2.13	0.43
	cNameV hen1 02 oZscore	.049	2.06	0.25	1.91	0.27
	cNameV myognf1 01 oZscore	.032	2.09	0.35	1.89	0.26
	cNameV ncx 01 oZscore	.032	2.62	0.54	1.99	0.31
	cNameV stat1 01 oZscore	.010	2.35	0.32	1.95	0.26
	cNameV tcf11mafg 01 oZscore	.017	2.27	0.29	1.97	0.26
Distributions of Illumina annotated genomic location		$P^a_{\text{chi square}}$	**Observed**	**Expected**	**Observed**	**Expected**
Location to Regulatory Feature	Promoter Associated	$6.1*10^{-04}$	1953	1846.0	1961	2068.0
	Unclassified Cell type specific	.026	462	498.1	594	557.9

a: Unadjusted p-value from Wilcoxon (epigraph webtool) or Chi Square test, only attributes with nominally significant ($p < 0.05$) differences are given