Epigenetic information loss is a common feature of multiple diseases and aging

Naor Sagy, Chieh Chang, Maayan Gal and Daniel Z Bar
The faculty of Medicine, Tel Aviv University, ISRAEL

Abstract

Aging is a major risk factor for a plethora of diseases. The information theory of aging posits that epigenetic information loss is a principal driver of the aging process. Despite this, the connection between epigenetic information loss and disease has not been thoroughly investigated. Here, we analyzed tissue-unique methylation patterns in healthy and diseased organs, revealing that for several diseases these patterns degrade, regressing to a mean form. We interpret this as epigenetic information loss, where tissue-unique patterns erode. Information loss is not limited to diseases. Age-related erosion of unique methylation patterns was observed in some tissues and cells, while other tissues and cells diverged away from the mean. Our findings demonstrate that analyzing methylation patterns in tissue-unique sites can effectively distinguish between patients and healthy controls across a range of diseases, and underscore the role of epigenetic information loss as a common feature in various pathological conditions.

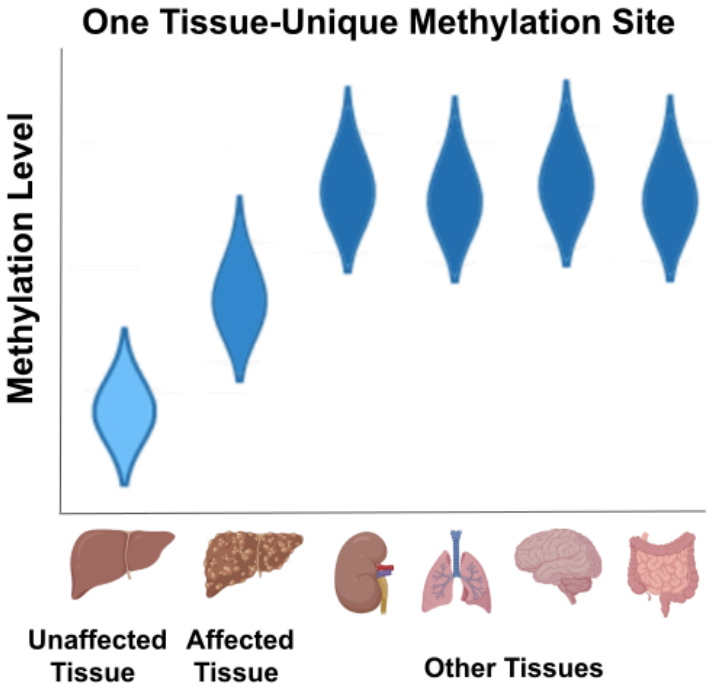


Figure 1: Schematic representation - Tissue unique methylation pattern regress toward the mean upon disease. *A single methylation site, showing low methylation in the liver and high in every other tissue, becomes more methylated in diseased livers.*

Introduction

Aging is the primary risk factor for numerous chronic diseases, such as cardiovascular diseases, cancer, and type 2 diabetes (T2D) [1]. Epigenetic alterations, including changes in methylation levels at specific genomic loci, are a well-known hallmark of aging [2–4]. Linear combinations of these loci, known as epigenetic clocks, strongly correlate with chronological age [5] and can predict all-cause mortality [6]. Behaviors that correlate with decreased healthspan and lifespan, including smoking and bad diet, accelerate some epigenetic clocks [6]. Stress may also induce rapid increase in biological age, as measured by epigenetic clocks [7]. Interestingly, recovery from stress may revert these changes [7]. Similarly, heterochronic parabiosis reduced the epigenetic age [8]. However, the causality of these methylation changes and why certain loci correlate with age remain unresolved [2].

The epigenetic theory of aging posits that the loss of epigenetic information disrupts youthful gene expression patterns, leading to cellular dysfunction and senescence [9–12]. Thus, information loss is suggested to drive the aging process [13]. Age-dependent loss of youthful gene expression patterns has been demonstrated in some cases [14–16], but not in others [17]. By contrast, single-cell level analysis of global coordination also suggests that transcriptional dysregulation is common or universal with aging [18].

Medical diagnosis typically relies on prior knowledge and the identification of relevant pathologies to explain symptoms. Specific DNA methylation patterns strongly correlate with medical conditions, but these correlations are only inferred by comparing methylation patterns of labeled cohorts [19–22]. In this study, we demonstrate that tissue-specific methylation patterns change with age and the progression of certain diseases. This pattern change allows us to predict Chronic Kidney Disease (CKD), liver diseases, T2D, and sun exposure in healthy skin using methylation data alone, without any additional information or prior knowledge. Moreover, in multiple cases, the direction of change is not random; instead, it moves towards a common form across all other tissues, which we interpret as epigenetic information loss. Our findings establish the loss of epigenetic information as a shared feature of multiple diseases (**Fig. 1**).

Results

Identification of tissue-specific methylation sites

Epigenetic information loss can take many forms, and can be challenging to identify and quantify in a consistent manner across multiple diseases, tissues and datasets. It was previously shown that in the kidney, methylation sites with a kidney-unique signature always regress towards the common form with CKD progression [23]. The unique methylation signature of the kidney, as manifested in these sites, erodes as the tissue becomes more similar to the rest of the body. The existence of tissue-unique methylation sites allows the exploration of such information loss, in a manner that is both tissue-specific and consistent across all examined tissues.

To facilitate this analysis, we generated a dataset of all tissue-specific methylation sites (**Fig. 2A-C**) based on the NGDC-CNCB dataset [24]. Of the 30 tissues across 5323 samples in the dataset, 23 had unique methylation sites, ranging from 55 in bladder to 50,872 in placenta (**Fig. 2D**; **Supp. Table 1**). In total, 87,922, or ~18% of sites tested, displayed a unique methylation pattern in at least one tissue, with over half of them in the placenta. These sites showed enrichment for known transcription-factor sites, with limited overlap between motifs identified in different tissues (**Supp. Fig. 1**).

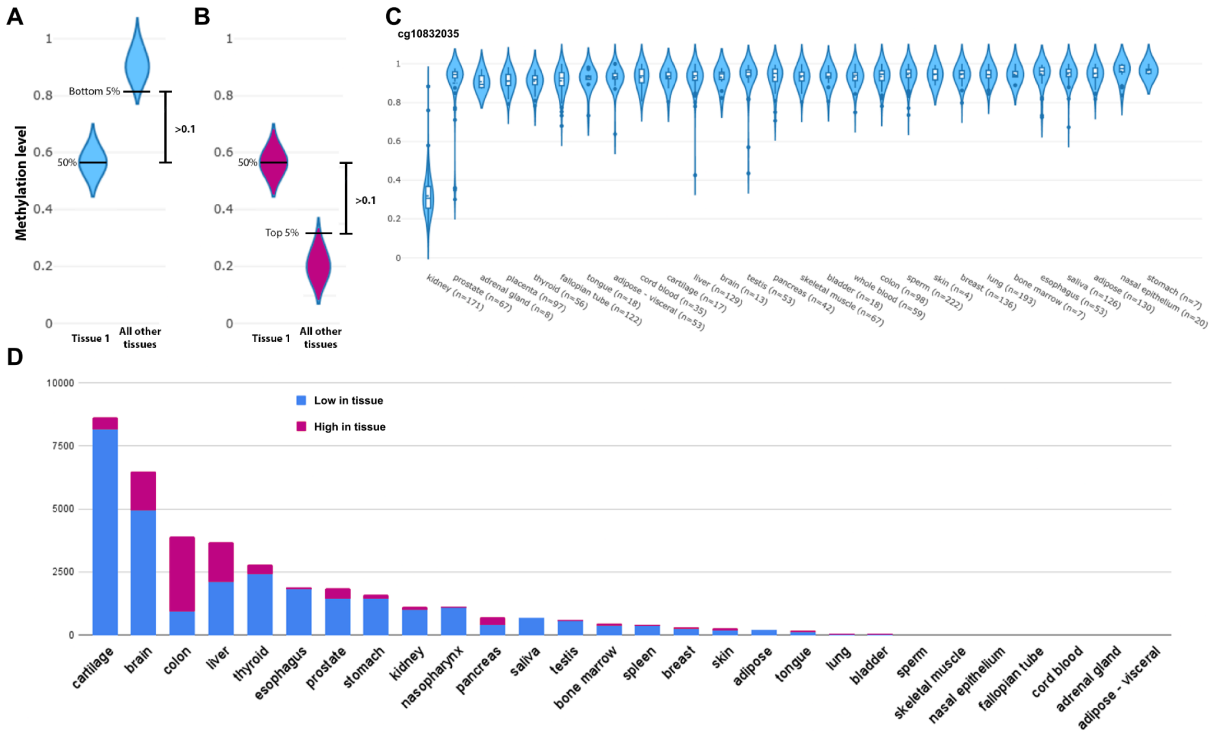


Figure 2: Identification of tissue-specific methylation sites. *Methylation sites that show a median absolute methylation value at least 0.1 below the bottom 5% quantile (**A**) or above the top 5% quantile (**B**) are considered unique. For example (**C**) shows a unique methylation pattern in the kidney. (**D**) Unique methylation sites for each of the tissues, excluding the placenta, that had over 40,000 low and 10,000 high sites.*

Hypothesis-free mapping of kidney-unique methylation sites correlates with CKD

Epigenetic information loss occurs in nearly all uniquely methylated sites as CKD progresses ([23] and **Supp. Tables 2-3**). To obtain an unbiased, global view of these data, we projected the methylation values of kidney sites (**Supp. Tables 1-3**) into a 2D space using principal component analysis (PCA). This mapping spatially separated individuals with high and low interstitial fibrosis (**Fig. 3A; Supp. Fig. 2A**). To determine if this is a general feature across all methylation sites, we repeated the analysis with randomly selected methylation sites (**Fig. 3B**) and extended the controls to sites with high (**Fig. 3C**) and low (**Fig. 3D**) kidney methylation variability across individuals. Among all cases, uniquely methylated sites demonstrated the strongest spatial separation. We concluded that uniquely methylated sites can be utilized for unbiased characterization of kidney functional decline.

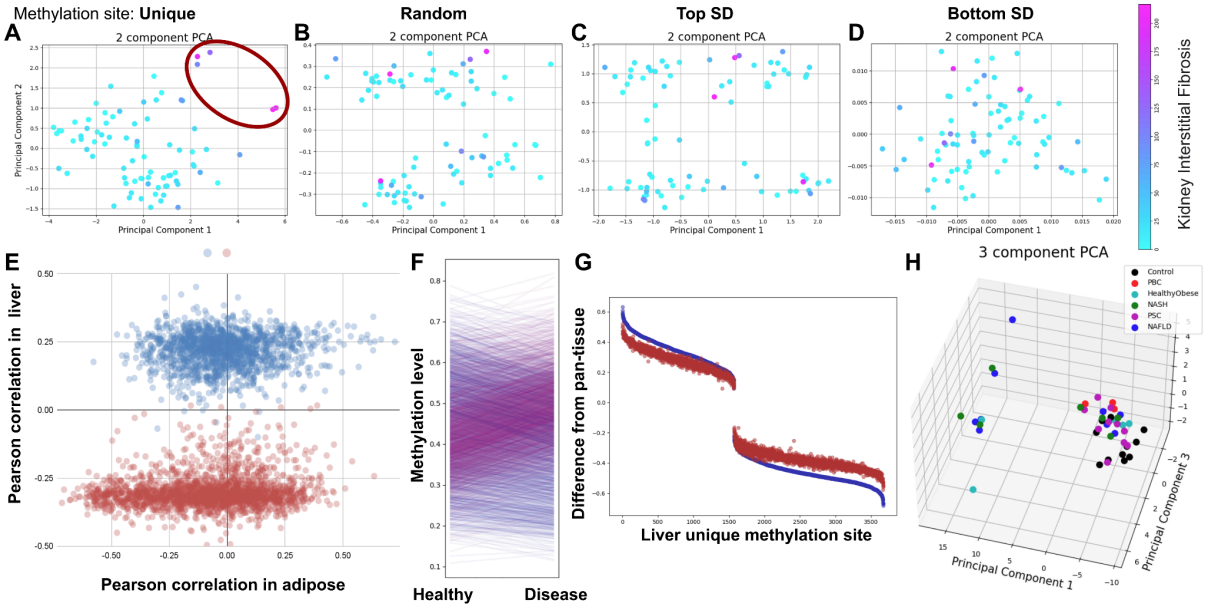


Figure 3: Epigenetic information loss in the kidney and liver. *PCA analysis of kidney-unique methylation sites clusters kidneys with high levels of interstitial fibrosis (**A**), but not when random (**B**), high variability (**C**) or low variability (**D**) sites are used. (**E**) Liver-unique methylation sites show a binomial distribution of correlations to liver diseases in liver samples (Y axis), but not in adipose (X axis). More methylated samples (red) show a negative correlation, while less methylated samples (blue) show a positive correlation. (**F**) Methylation levels (mean) for sites that are more (red) and less (blue) methylated in the liver (compared to the rest of the body) reveals regression to the mean in disease. (**G**) Differences between each liver-unique methylation site mean and pan-tissue average, in healthy (blue) and sick (red) samples. Regression to the mean (red closer to zero) is observed in nearly all sites. (**H**) PCA of liver-unique methylation sites in men only.*

Multiple liver diseases show loss of epigenetic information

The liver is renowned for its exceptional regeneration capacity, which may make it less susceptible to aging effects compared to other tissues. Indeed, the risk of some liver diseases does not appear to correlate with age, at least in men [25,26]. However, liver function declines with age, and aging is associated with increased severity and poor prognosis of multiple liver diseases [27]. To explore the role of epigenetic information loss in liver diseases, we analyzed a combined multi-disease and multi-tissue dataset ([28]; GSE61256; N = 137; **Supp. Tables 2-3**). It includes, in addition to liver methylation samples from healthy individuals, data for healthy but obese participants, as well as Primary Biliary Cirrhosis (PBC), Non-Alcoholic SteatoHepatitis (NASH), Primary Sclerosing Cholangitis (PSC) and Non-Alcoholic Fatty Liver Disease (NAFLD) patients. Methylation data for adipose and muscle tissues were also available for some donors. Our tissue-specific dataset contains 3,673 liver-specific methylation sites, all of which were found in this dataset, enabling high-confidence detection of epigenetic information loss. The correlation between liver methylation in the NGDC-CNCB dataset, used to derive the unique sites, and the controls of GSE61256, was r = 0.99 (**Supp. Fig. 3**), indicating high technical reproducibility across different datasets. We first explored whether methylation levels in these sites correlate with all diseases. Remarkably, liver-unique methylation sites exhibited a bimodal distribution with a positive and a negative peak, and almost no sites showed low correlation to liver disease (**Fig. 3E**, Y-axis). In contrast, the same sites in the same patients, but in adipose tissue, showed a normal distribution (**Fig. 3E**, X-axis). We investigated what

distinguishes sites that positively vs. negatively correlate with liver disease. Overwhelmingly, these two populations were distinguished by their methylation status in the liver compared to the rest of the body (3,658 of 3,673; $p < 10^{-1000}$). Thus, 99.6% of these sites regressed to the mean with liver disease progression (**Fig. 3F-G**).

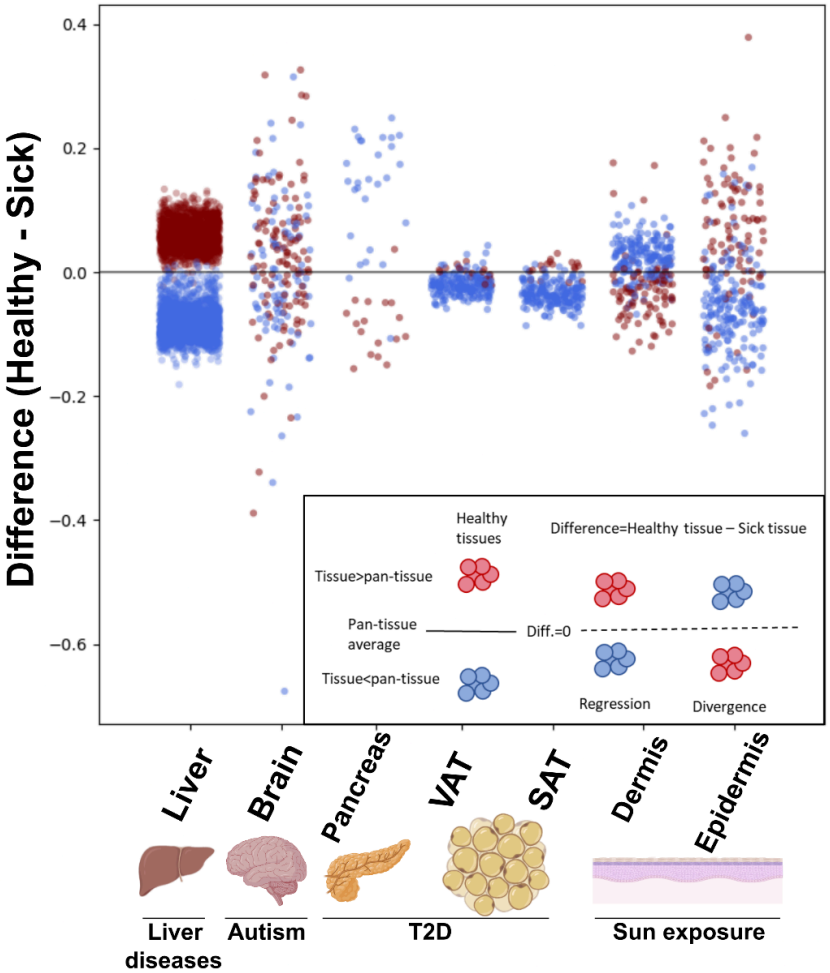


Figure 4: Methylation difference in disease across seven tissues. Inset *- schematic explanation: Red dots represent sites methylated above other tissues, while blue dots represent lower methylation levels. When looking at the difference between healthy and sick, red above zero and blue below indicate regression to the mean, while the opposite indicates divergence. Y-axis: Difference in methylation levels, for each tissue-unique methylation site, between healthy and sick (cohort averages). Differences shown for liver (liver diseases), Brain (autism), pancreas (T2D), VAT and SAT (T2D) and dermis and epidermis (sun exposure).*

Contrary to our expectations based on these results, PCA of these samples showed a poor separation between healthy and diseased livers (**Supp. Fig. 4**). We hypothesized that another source of variance was dominating this analysis. Indeed, we found that men and women formed two distinct clusters when performing PCA using high-variability sites (**Supp. Fig. 5**), in line with previous findings that the human liver has a sex-specific methylation profile in autosomes [29]. When analyzing men separately, controls clustered together and partially separated from diseases (**Fig. 3H**). Control methylation sites exhibited weaker separation of the control group (**Supp. Fig. 6**). Surprisingly, analyzing women only yielded a relatively poor separation of controls (**Supp. Fig. 7**). That can be attributed to gene expression differences between men and women [30–32]. These inherent dissimilarities might manifest in methylation profiles and different hormone levels in the liver [33], and the nature of PCA analysis which segregates samples according to maximum variance and not by a certain biological orientation. These results suggest that liver disease results in epigenetic information loss in the liver.

Type 2 diabetes shows epigenetic information loss in adipose but not in the pancreas

T2D is the most prevalent form of diabetes, characterized by high blood sugar, insulin resistance, a relative lack of insulin, and changes to the pancreas [34–36]. Age is a major risk factor for T2D, and it correlates with increased biological age as indicated by epigenetic clocks [37]. To investigate if epigenetic information is lost in the pancreas of T2D patients, we applied our analysis to a dataset mapping 27,578 CpG sites in human pancreas samples from T2D patients and controls ([38]; GSE21232; N = 16). Of the 705 pancreas-unique methylation sites, only 45 were covered by the methylation array used. Among these, 28 were less methylated in the pancreas. Contrary to our findings in liver and kidney, 26 of the 28 sites diverged away from the mean in T2D (**Fig. 4**; $p < 10^{-5}$). Of the 17 sites that were more methylated in the pancreas, divergence was observed in 16 (**Fig. 4**; $p < 10^{-3}$). PCA almost completely separated diabetic from non-diabetic pancreas samples (**Supp. Fig. 8**). Interestingly, using random sites or sites with high variance in the pancreas — but not sites of low variance — also separated diabetic from non-diabetic pancreases (**Supp. Fig. 8**). Thus, the epigenetic changes are robust enough for T2D to be detected in multiple sites.

Diabetes is closely linked to obesity and the accumulation of adipose tissue. However, not all adipose tissues confer the same risk. Specifically, visceral adipose (VAT) poses a significantly greater risk than subcutaneous fat (SAT) [39]. To test if epigenetic information is lost in diabetes, we examined a dataset of VAT and SAT from T2D patients and controls ([40]; GSE162166; N = 40). Globally, some methylation patterns were observed. In VAT, 84% of all methylation sites showed elevated methylation in T2D patients. Furthermore, examining methylation sites with a robust change in any direction—measured by an absolute correlation greater than 0.3—94% of sites showed increased methylation. Interestingly, this was observed in VAT only, whereas in SAT, only 46% showed elevated methylation. Adipose tissue has 222 unique sites (**Fig. 2D**, **Supp. Tables 2-3**). Among these, 212 were covered by the methylation profiling dataset, including VAT and SAT from T2D patients and controls ([40]; GSE162166). Methylation patterns regressed towards the mean at 90% of these sites in VAT and 95% in SAT (**Fig. 4** - blue dots below the zero line and red dots above it). Although statistically significant, this finding is not particularly surprising, as 91% of the unique sites exhibited lower methylation levels in adipose tissue, implying that the general trend in VAT can largely account for this observation. Considering the remaining 9% (N = 18) of the sites, regression towards the mean (and contrary to the 84% trend observed) was evident in 9 of them ($p = 10^{-3}$). In SAT, 14 of the 18 sites regressed towards the mean ($p < 10^{-7}$). Consequently, we concluded that a global increase in methylation is a characteristic feature of VAT, but not SAT, in T2D patients, and both tissues experience a loss of epigenetic information. However, these findings were not accompanied by clear separation by PCA, as healthy and T2D patients did not form distinct clusters in neither VAT nor SAT, when using adipose-unique sites (**Supp. Fig. 9**).

Autism does not show significant epigenetic information loss

Autism is a neurodevelopmental disorder characterized by difficulties with social interaction and communication. Symptoms of autism typically appear early in childhood, and it has not been linked to aging. We tested if epigenetic information is lost in autism. The brain contains a relatively large number of unique methylation sites: 6,491. Of these, 209 were covered by a dataset of brain methylation ([41]; GSE38608; N = 36). In autism, we observed information loss in 125 of these. While these findings were marginally statistically significant ($p = 0.005$), they appeared very weak compared to kidney and liver diseases (**Fig. 4** - no clear separation of red and blue above or below

the zero line in brain, unlike liver). We performed PCA using all brain-unique methylation sites, and autistic brains appeared indistinguishable from controls (**Supp. Fig. 10**). A clear signature was observed, but it was not related to autism (see "Epigenetic information loss during aging" below). Together, these results suggest that loss of tissue-specific methylation information is not a universal feature of all diseases and disorders. However, these results do not exclude the possibility that a different analysis or sample correction may reveal information loss.

Epigenetic information is lost in environmentally stressed tissue

The skin serves as a barrier that protects the body from external insults, such as UV radiation from the sun. Sun exposure prematurely ages the skin, termed photoaging [42]. Photoaging manifests differently in the dermis and epidermis, which can be partly attributed to the limited penetration of UV-B into the dermis [43]. We investigated whether typical environmental stressors to tissue induce epigenetic information loss. To do so, we compared methylation data from sun-exposed and sun-protected skin samples of both dermis and epidermis, in young and old individuals [44]. Of the 277 skin unique methylation sites, 275 were present on the dataset (GSE51954; N = 78). In the dermis, divergence from the mean upon sun exposure was observed (**Fig. 4** - dermis. Blue dots above zero, red below; **Supp. Table 3**). In contrast, in the epidermis, regression to the mean occurred at 223 and 226 sites for young and old individuals, respectively (**Fig. 4**; **Supp. Table 3**). When mapping methylation-levels of skin-unique methylation sites into 3D space using PCA, dermis and epidermis samples were clearly separated (**Fig. 5A, Supp. Fig. 2B**). However, samples from older versus younger donors exhibited only partial separation (**Fig. 5A**). Sun-exposed and sun-protected samples from each tissue type clustered separately, with the expected stronger separation observed in the epidermis. Analyzing the same samples using methylation sites that were not tissue-specific, including random sites and sites with high standard deviation, did not yield such a distinct separation (**Fig. 5B-D**). Consequently, we conclude that sun exposure is associated with epigenetic information loss in the epidermis.

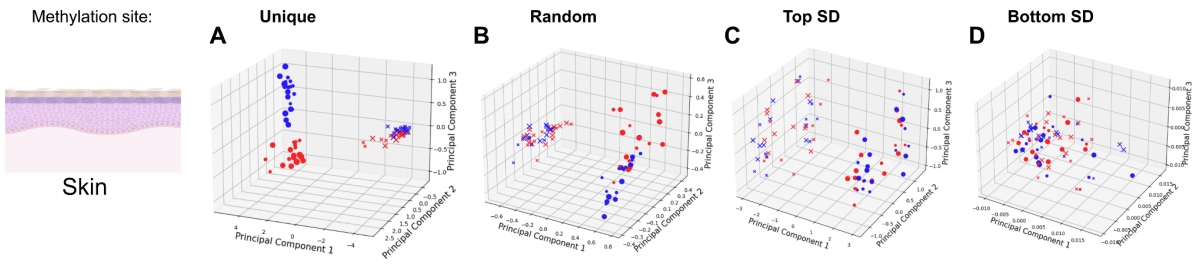


Figure 5: Separation based on tissue-unique sites following environmental stress. *3D PCA analysis of skin-unique methylation sites separated sun-exposed (red) from sun-protected (blue) samples (**A**), but not when random (**B**), high variability (**C**) or low variability (**D**) sites are used. X - dermis; o - epidermis; size indicates young (small) or old (large).*

Epigenetic information loss during aging

The epigenetic theory of aging posits that the loss of epigenetic information contributes to the aging process ([10,13,45,46]). Tissue functionality declines with age, similar to the deterioration observed in multiple diseases. This decline may vary among tissues, as some experience a faster loss of function on average than others. Nonetheless, a strong epigenetic signature of aging has been observed in most tissues [5]. We tested if tissue-unique epigenetic information is lost during aging, focusing on

three tissues. The brain has low regenerative capacity and slow cell turnover. With aging, cognitive functions decline and memory formation may become impaired. In contrast, the intestine exhibits high regenerative capacity and cell turnover. Aging in the intestine is characterized by reduced self-renewal capacity of intestinal stem cells and diminished tissue self-repair functionality [47]. When plotting the brain-unique methylation sites for autistic and nonautistic individuals, we noticed PCA separated the samples by age (**Supp. Fig. 11** and **Fig. 6A**). Indeed, PC1 strongly correlated with the age of the individuals (**Fig. 6A, Supp. Fig. 2C**). However, random sites, as well as sites with high variance also produced the same result (**Fig. 6B-D**). Of the 209 brain-unique sites covered by the dataset, regression to the mean was observed in 148, and not observed in 61 sites ($p < 10^{-9}$). Next we analyzed normal colon mucosa samples ([48]; GSE48988; N = 178). Of the 3894 intestine-unique sites, 290 were found in the dataset. Of these, 87% (252) had higher methylation levels in the intestine than in other tissues. 48 sites did not produce any methylation values and were discarded. Divergence from the mean was observed in 237 of the 242 remaining sites, while regression was observed in just 5 sites ($p < 10^{-63}$; **Fig. 6J**). PCA generated a clear age gradient when using intestine-unique sites, but not in any of the controls (**Fig. 6F-I, Supp. Fig. 2D**). We concluded that the tissue-unique epigenetic signature degrades with age, but only in some tissues. This result is independent of whether unbiased PCA clustering separates samples by age.

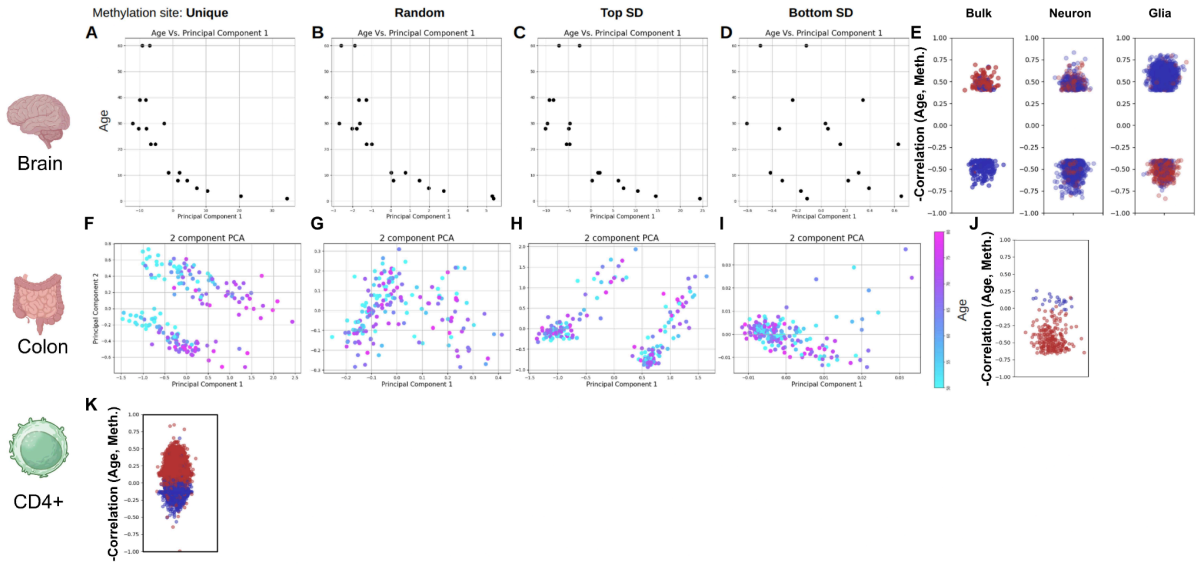


Figure 6: Epigenetic information loss in aging. *PC1 vs age in brain samples, using (**A**) unique (**B**) random (**C**) high variability and (**D**) low variability sites. (**E**) Correlation of methylation levels and age for brain-unique methylation site, in bulk, neuronal and glial cells. PCA of intestine samples, using (**F**) unique (**G**) random (**H**) high variability and (**I**) low variability sites. Color indicates age.(**J**) Correlation of methylation levels and age for intestine-unique methylation site (**K**) Correlation of methylation levels and age for intestine-unique methylation site. All correlations are plotted inversely for schematic consistency (red above zero for regression).*

Epigenetic information is lost in some, but not all, cell types

To investigate whether epigenetic information loss occurs within a cell population, we analyzed a dataset of human brain samples ([49]; GSE41826; N = 145) which included bulk tissue samples as well as fluorescence-activated cell sorted (FACS) neuronal and non-neuronal (predominantly glia) cells. In the bulk samples, consistent with our previous observation, regression to the mean with age was observed in 85% (1302/1531; $p < 10^{-181}$) of more methylated sites and 77% (3819/4926; $p < 10^{-344}$) of

less methylated sites. Analyzing only sites that robustly correlate with age (R > 0.4) increased these numbers to 96% ($p < 10^{-31}$; **Fig. 6E**) and 86% ($p < 10^{-30}$) respectively. This trend was mimicked in neurons, where regression to the mean was observed in 71% ($p < 10^{-9}$) and 70% ($p < 10^{-32}$), of more and less methylated sites, respectively. Surprisingly, the opposite trend was observed in glia cells, where methylation levels predominantly diverged away from the mean (85%; $p < 10^{-49}$ and 94%; $p < 10^{-390}$, respectively).

To further corroborate these findings, we analyzed the NGDC-CNCB blood DNA methylation dataset [50]. This independent dataset comprises 25 types of blood samples (N=3402), including mixed and cell-type-specific samples. We focused on purified naive CD4+ T cells (CD4+). As these cells were not present in our original analysis, we identified 29,137 methylation sites unique to CD4+ cells (**Supp. Table 4**), compared to the remaining 24 blood sample types. Correlation of methylation level and age was calculated for each unique site for samples where age was available (N=71). The majority of the sites were more methylated in CD4+ compared to the remaining samples. Of these, 96% (20825/21556; $p < 10^{-1000}$) of sites regressed towards the mean with age (Fig. 6K - red dots) . Similarly, of the less methylated sites 93% (7102/7581; $p < 10^{-1000}$) regressed towards the mean (Fig.6K - blue dots). Thus, some cell types regress to the mean, while others diverge away from it.

Methylation-based disease classification

Tissue-specific methylation signature changes in various diseases raise the question of whether these unique sites can be used for disease classification. We applied three classification models using only methylation data. We were able to distinguish between low and high GFR (kidney) with an accuracy as high as 82.35%, between sun-exposed and sun-protected skin samples in the dermis with an accuracy of 80%, and in the epidermis with an accuracy of 100%, respectively (**Supp. Fig. 11**)

Discussion

Aging is accompanied by epigenetic changes. By focusing on DNA methylation sites with unique signatures compared to other tissues, we could isolate genomic loci where age-dependent changes in methylation can be interpreted. In these sites, observing the direction of change can inform us if it is random, hypomethylation, hypermethylation, divergence or regression to the mean. In multiple instances, including diseases, environmental stress, and natural aging, these sites predominantly changed by regression to the mean. For example, 99.6% of liver-unique sites regressed to the mean in liver diseases. We interpret these results as a loss of epigenetic information. The unique methylation signature that characterizes the tissue becomes more similar to other tissues in the body, akin to the smoothening of the epigenetic landscape or exdifferentiation reported elsewhere [11,13,56]. These results do not imply that a similar process does not occur in other methylation sites. On the contrary, the fact that such a large fraction of the a-priori selected sites showed this trend suggests this is likely to be a more general phenomenon. Moreover, other types of epigenetic information may exhibit a similar phenomenon. However, analyzing these sites will require the development of expectations as to the direction of change, to distinguish information loss from changes that may reflect different processes. Interestingly, some tissues that do not show this phenomenon demonstrate the opposite effect. Methylation sites diverge away from the common form as the tissue becomes more distinct. Thus, our experiments do not support a role of epigenetic information loss, at least for these sites and bulk tissue. We postulate that another unknown

mechanism, actively counteracting the smoothening of the epigenetic landscape, is dominant in these samples.

One possible explanation for the loss of epigenetic information in disease and aging is changes to tissue cell composition. Such changes, including replacement of resident-tissue cells by fibroblasts and accumulation of immune cells, have been demonstrated in aging and disease in some tissues, but not in others [57–60]. A non-exclusive alternative is changes within the cell population. For example, in epithelial to mesenchymal transition (EMT), epithelial cells undergo changes that enable them to assume a mesenchymal cell phenotype [61]. EMT has been associated with fibrosis in kidney, liver, and intestine [61–65] and changes in DNA methylation have been suggested to causally underlie EMT [66]. In human brain samples, regression to the mean with age was observed in brain-unique sites of bulk tissue. When looking at sorted cells, we observed regression to the mean in neurons, but not glial cells of aging brains. In glial cells methylation levels diverged away from the mean, and the tissue-unique signature appears to become stronger. Similarly, regression to the mean with age was observed in unique sites in CD4+ cells. Future analysis of samples separated by cell type will enable us to determine the relative contribution of each of these mechanisms to epigenetic information loss, as these may differ across organs and pathologies.

Epigenetic information holds diagnostic value. Using the same algorithms and solely relying on methylation data, we were able to classify diseased and healthy organs at high rates. These results highlight the diagnostic potential of methylation data. As some disease induced changes to the methylome are reversible [7], and heterochronic parabiosis can reverse or delay some changes to epigenetic age [8], it would be interesting to investigate if treatment may reverse the regression to the mean in affected organs. Moreover, it has the potential to suggest which organs are affected by a medical condition. More sophisticated analyses may even distinguish different diseases in the same organ and help classify new disease subtypes. Finally, the epigenetic signature of these tissue-specific sites may allow for their identification from cell-free DNA in blood.

Major limitations of this work include predominantly bulk sample analysis, with limited cell-type and no single-cell data, and lack of longitudinal data. We focused on tissue-level methylation levels across multiple sites. However, each tissue is composed of multiple cell types, often arranged in functionally-important spatial patterns. The bulk analysis limits the extractable information. Cell-type unique methylation sites may have remained undiscovered in our analysis as the signal was suppressed by other cells of that tissue. Similarly, changes that manifest only in a subset of cells may not be detected. For example, beta cells are lost in T2D. However, as they comprise only 1-2% of the pancreas mass, their loss in T2D is unlikely to be captured by this analysis. Thus, these experiments do not inform us if epigenetic information loss plays a role in beta cell loss in T2D. The addition of such cell-specific information, both for unique methylation site identification and for deeper tissue analysis, will enable greater sensitivity while informing us which cells are losing information and by which mechanism. Next, while we were able to demonstrate that methylation-based disease classification is possible, our study focused on comparing disease samples to control samples. Future studies should aim to identify changes in methylation patterns within individuals over time, as this would enable us to identify epigenetic information loss in an individual rather than in comparison to a control. Finally, we note that while aggregation of hundreds to thousands of sites across tens of

samples enabled robust statistics, the average data content per individual site and patient is low. This is likely the reflection of the stochastic nature of this process.

Methods

Unique methylation sites

A site was considered unique for a tissue if the average methylation value in that tissue was either (i) at least 0.1 higher than the 95% quantile of all other tissues (combined) or (ii) at least 0.1 lower than the 5% quantile of all other tissues in the NGDC-CNCB dataset. CD4+ unique sites were identified by applying the same analysis to the NGDC-CNCB blood dataset. Methylation sites with fewer than 2,000 valid values in the dataset were omitted from the analysis. The dataset, containing Illumina HumanMethylation 450 BeadChip probes, was screened for the above criteria using a custom Python script. The results were manually verified for a random subset. Among the 30 available tissues, 8 did not have unique sites. It should be noted that these criteria do not entirely eliminate the possibility of two tissues sharing the same unique site. Parsing the dataset, analyzing the data and sorting the unique sites was done using Python.

Datasets used

Methylation data were obtained from the Gene Expression Omnibus [1,68]. All datasets used are listed in **Supplementary Table 2**. Analysis was conducted on the processed Series Matrix Files, which were separated into a methylation data file and an experiment metadata file.

Motif enrichment

We analyzed the unique sites for de novo and known motifs across all tissues using Hypergeometric Optimization of Motif EnRichment (HOMER) [69,70]. We extracted ±500bp around the start of methylation probes using the Python Bio.Entrez package [71], and obtained chromosome, strands, and genomic coordinate data from the HM450 hg19 database. Sequences were obtained for all unique methylation sites across all tissues and for 3,000 random methylation sites, which were used as background for motif enrichment detection. Sequences were saved in Fasta format using Python. HOMER was run for each tissue's unique sites separately, against the background sequences, and the motif results were filtered by p-value $< 10^{-3}$ and occurrence percentage in the background <10% using a Python script. Enrichment was calculated as the log2 of the percentage in the target divided by the percentage in the background.

Principal component analysis

PCA linear dimensionality reduction was employed to project n-dimensional data (n CpG sites) onto 2D/3D space. We utilized the Python scikit-learn package to perform PCA. The data were parsed and rearranged from a dataframe format to comply with the package's format requirements, representing probes as dimensions. Whenever a subset of the database was required, we either reconstructed the input files to include only the required dataset or filtered out test subjects in the code prior to implementing PCA on the data. Filtering was conducted on metadata stored in a separate dataframe and/or on methylation sites.

Disease classification

Values for unique methylation sites were extracted from the databases for each relevant tissue analyzed in the respective datasets and projected onto 2D using PCA. Samples were classified using three models: 1) Ratio of average distances from the sample to healthy and sick samples. 2) Comparing distance from the sample to the control group center of mass and distance from the sample to the disease group center of mass. 3) If the distance from a sample in 2D space was less than twice the standard deviation from the mean location of the disease group, the sample was classified as sick.

To estimate the accuracy of each model, in each iteration, one sample was omitted from the PCA analysis and then transformed and projected onto 2D space using the PCA weights vector. The three models were then applied to each sample, and accuracy was calculated by comparing classification with true results from the database.

All the models projected a single sample onto an existing PCA space based on the rest of the samples, and subsequently classified it into one of the groups without prior knowledge.

Statistics

P-values were calculated using the binomial distribution in R (dbinom), except when otherwise specified. Where applicable, technical duplicates were removed from data before the analysis.

Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** Generated data are available as supplementary files.

References

1. Niccoli, T. & Partridge, L. Ageing as a risk factor for disease. *Curr. Biol.* **22**, R741–52 (2012).
2. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. *Cell* (2022) doi:10.1016/j.cell.2022.11.001.
3. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
4. Issa, J.-P. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124**, 24–29 (2014).
5. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
6. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
7. Poganik, J. R. *et al.* Biological age is increased by stress and restored upon recovery. *Cell Metab.* **35**, 807–820.e5 (2023).
8. Zhang, B. *et al.* Multi-omic rejuvenation and life span extension on exposure to youthful

circulation. *Nat Aging* (2023) doi:10.1038/s43587-023-00451-9.

9. Lu, Y. *et al.* Reprogramming to recover youthful epigenetic information and restore vision. *Nature* **588**, 124–129 (2020).
10. Yang, J.-H. *et al.* Erosion of the Epigenetic Landscape and Loss of Cellular Identity as a Cause of Aging in Mammals. *bioRxiv* 808642 (2019) doi:10.1101/808642.
11. Sinclair, D. A. & LaPlante, M. D. *Lifespan: Why We Age—and Why We Don't Have To*. (Simon and Schuster, 2019).
12. Imai, S. & Kitano, H. Heterochromatin islands and their dynamic reorganization: a hypothesis for three distinctive features of cellular aging. *Exp. Gerontol.* **33**, 555–570 (1998).
13. Yang, J.-H. *et al.* Loss of epigenetic information as a cause of mammalian aging. *Cell* (2023) doi:10.1016/j.cell.2022.12.027.
14. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* **441**, 1011–1014 (2006).
15. Busuttil, R., Bahar, R. & Vijg, J. Genome dynamics and transcriptional deregulation in aging. *Neuroscience* **145**, 1341–1347 (2007).
16. Martinez-Jimenez, C. P. *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
17. Warren, L. A. *et al.* Transcriptional instability is not a universal attribute of aging. *Aging Cell* **6**, 775–782 (2007).
18. Levy, O. *et al.* Age-related loss of gene-to-gene transcriptional coordination among single cells. *Nat Metab* **2**, 1305–1315 (2020).
19. Dritsoula, A. *et al.* 'Epigenome-wide methylation profile of chronic kidney disease-derived arterial DNA uncovers novel pathways in disease-associated cardiovascular pathology.' *Epigenetics* **16**, 718–728 (2021).
20. Galbraith, K. & Snuderl, M. DNA methylation as a diagnostic tool. *Acta Neuropathol Commun* **10**, 71 (2022).

21. Cerrato, F. *et al.* DNA Methylation in the Diagnosis of Monogenic Diseases. *Genes* **11**, (2020).

22. Sadikovic, B. *et al.* Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet. Med.* **23**, 1065–1074 (2021).

23. Sagy, N., Meyrom, N., Beckerman, P., Pleniceanu, O. & Bar, D. Z. Kidney-specific methylation patterns correlate with kidney function and are lost upon kidney disease progression. *bioRxiv* 2022.09.19.508466 (2022) doi:10.1101/2022.09.19.508466.

24. Xiong, Z. *et al.* EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. *Nucleic Acids Res.* **50**, D1004–D1009 (2022).

25. Jeong, E. H. *et al.* Regional prevalence of non-alcoholic fatty liver disease in Seoul and Gyeonggi-do, Korea. *Clin. Mol. Hepatol.* **19**, 266–272 (2013).

26. Amarapurkar, D. *et al.* Prevalence of non-alcoholic fatty liver disease: population based study. *Ann. Hepatol.* **6**, 161–163 (2007).

27. Kim, I. H., Kisseleva, T. & Brenner, D. A. Aging and liver disease. *Curr. Opin. Gastroenterol.* **31**, 184–191 (2015).

28. Horvath, S. *et al.* Obesity accelerates epigenetic aging of human liver. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15538–15543 (2014).

29. García-Calzón, S., Perfilyev, A., de Mello, V. D., Pihlajamäki, J. & Ling, C. Sex Differences in the Methylome and Transcriptome of the Human Liver and Circulating HDL-Cholesterol Levels. *J. Clin. Endocrinol. Metab.* **103**, 4395–4408 (2018).

30. Zhang, Y. *et al.* Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. *PLoS One* **6**, e23506 (2011).

31. Lau-Corona, D., Bae, W. K., Hennighausen, L. & Waxman, D. J. Sex-biased genetic programs in liver metabolism and liver fibrosis are controlled by EZH1 and EZH2. *PLoS Genet.* **16**, e1008796 (2020).

32. Clodfelter, K. H. *et al.* Sex-dependent liver gene expression is extensive and largely dependent upon signal transducer and activator of transcription 5b (STAT5b): STAT5b-dependent activation

of male genes and repression of female genes revealed by microarray analysis. *Mol. Endocrinol.* **20**, 1333–1351 (2006).

33. Kasarinaite, A., Sinton, M., Saunders, P. T. K. & Hay, D. C. The Influence of Sex Hormones in Liver Function and Disease. *Cells* **12**, (2023).
34. Bogardus, C., Lillioja, S., Howard, B. V., Reaven, G. & Mott, D. Relationships between insulin secretion, insulin action, and fasting plasma glucose concentration in nondiabetic and noninsulin-dependent diabetic subjects. *J. Clin. Invest.* **74**, 1238–1246 (1984).
35. Reaven, G. M. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes* **37**, 1595–1607 (1988).
36. Macauley, M., Percival, K., Thelwall, P. E., Hollingsworth, K. G. & Taylor, R. Altered volume, morphology and composition of the pancreas in type 2 diabetes. *PLoS One* **10**, e0126825 (2015).
37. Bahour, N. *et al.* Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers. *Geroscience* **44**, 415–427 (2022).
38. Volkmar, M. *et al.* DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J.* **31**, 1405–1426 (2012).
39. Bergman, R. N. *et al.* Why visceral fat is bad: mechanisms of the metabolic syndrome. *Obesity* **14 Suppl 1**, 16S–19S (2006).
40. Jin, N. *et al.* Integratome analysis of adipose tissues reveals abnormal epigenetic regulation of adipogenesis, inflammation, and insulin signaling in obese individuals with type 2 diabetes. *Clin. Transl. Med.* **11**, e596 (2021).
41. Ginsberg, M. R., Rubin, R. A., Falcone, T., Ting, A. H. & Natowicz, M. R. Brain transcriptional and epigenetic associations with autism. *PLoS One* **7**, e44736 (2012).
42. Rittié, L. & Fisher, G. J. Natural and sun-induced aging of human skin. *Cold Spring Harb. Perspect. Med.* **5**, a015370 (2015).
43. Debacq-Chainiaux, F., Leduc, C., Verbeke, A. & Toussaint, O. UV, stress and aging. *Dermatoendocrinol.* **4**, 236–240 (2012).

44. Vandiver, A. R. *et al.* Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biol.* **16**, 80 (2015).

45. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

46. Muñoz-Najar, U. & Sedivy, J. M. Epigenetic control of aging. *Antioxid. Redox Signal.* **14**, 241–259 (2011).

47. Funk, M. C., Zhou, J. & Boutros, M. Ageing, metabolism and the intestine. *EMBO Rep.* **21**, e50047 (2020).

48. Noreen, F. *et al.* Modulation of age- and cancer-associated DNA methylation change in the healthy colon by aspirin and lifestyle. *J. Natl. Cancer Inst.* **106**, (2014).

49. Guintivano, J., Aryee, M. J. & Kaminsky, Z. A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290–302 (2013).

50. Download. https://ngdc.cncb.ac.cn/ewas/datahub/download.

51. Zhang, Z. *et al.* Hierarchical deconvolution for extensive cell type resolution in the human brain using DNA methylation. *Front. Neurosci.* **17**, 1198243 (2023).

52. Zheng, S. C. *et al.* EpiDISH web server: Epigenetic Dissection of Intra-Sample-Heterogeneity with online GUI. *Bioinformatics* **36**, 1950–1951 (2019).

53. Cai, M., Zhou, J., McKennan, C. & Wang, J. scMD: cell type deconvolution using single-cell DNA methylation references. *bioRxiv* (2023) doi:10.1101/2023.08.03.551733.

54. Zhu, T. & Teschendorff, A. E. Cell-Type Deconvolution of Bulk DNA Methylation Data with EpiSCORE. *Methods Mol. Biol.* **2629**, 23–42 (2023).

55. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).

56. Cutler, R. G. The dysdifferentiative hypothesis of mammalian aging and longevity. *The aging brain* **20**, 1–18 (1982).

57. Fu, Y., Yu, Y., Paxinos, G., Watson, C. & Rusznák, Z. Aging-dependent changes in the cellular composition of the mouse brain and spinal cord. *Neuroscience* **290**, 406–420 (2015).

58. Wang, Q. *et al.* The Aged Intestine: Performance and Rejuvenation. *Aging Dis.* **12**, 1693–1712 (2021).

59. Lee, H., Hong, Y. & Kim, M. Structural and Functional Changes and Possible Molecular Mechanisms in Aged Skin. *Int. J. Mol. Sci.* **22**, (2021).

60. Csekes, E. & Račková, L. Skin Aging, Cellular Senescence and Natural Polyphenols. *Int. J. Mol. Sci.* **22**, (2021).

61. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).

62. Roeder, S. S. *et al.* Changes in glomerular parietal epithelial cells in mouse kidneys with advanced age. *Am. J. Physiol. Renal Physiol.* **309**, F164–78 (2015).

63. Zeisberg, M. *et al.* Fibroblasts derive from hepatocytes in liver fibrosis via epithelial to mesenchymal transition. *J. Biol. Chem.* **282**, 23337–23347 (2007).

64. Kim, K. K. *et al.* Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 13180–13185 (2006).

65. Bechtel, W. *et al.* Methylation determines fibroblast activation and fibrogenesis in the kidney. *Nat. Med.* **16**, 544–550 (2010).

66. Galle, E. *et al.* DNA methylation-driven EMT is a common mechanism of resistance to various therapeutic agents in cancer. *Clin. Epigenetics* **12**, 27 (2020).

67. Brown, T. M. & Krishnamurthy, K. Histology, Dermis. in *StatPearls* (StatPearls Publishing, 2022).

68. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991–5 (2013).

69. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* **29**, 1836–1846 (2019).

70. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

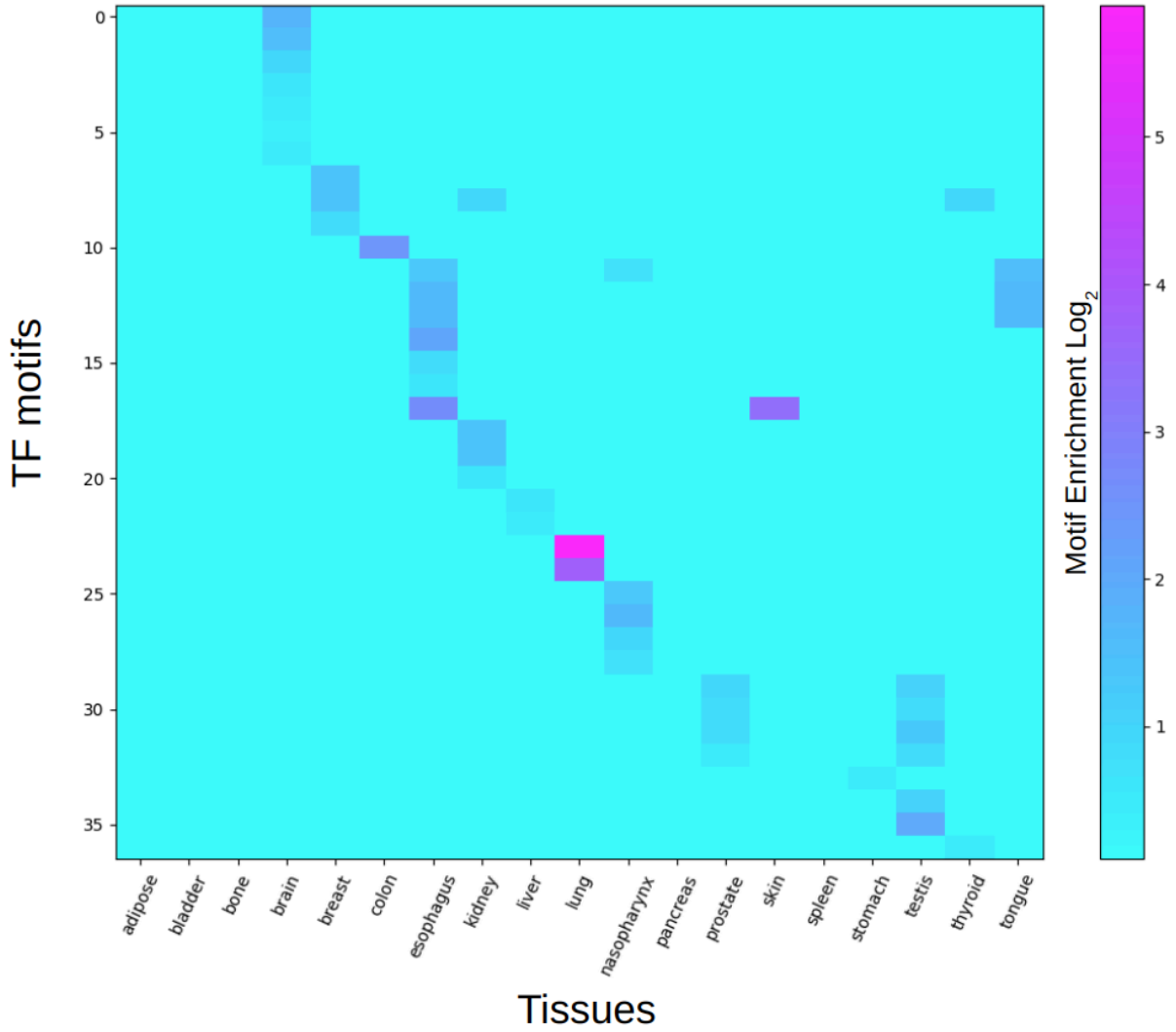
71. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

Acknowledgements
We thank Professors Yosef Gruenbaum, Yuval Ebenstein, Lihi Adler-Abramovich, and Maayan Gal, as well as the Bar lab members, for their critical review and suggestions.
Funding: The work was supported by the Israeli Science Foundation (grants 654/20 and 632/20 to DZB) and the Center for Artificial Intelligence & Data Science in Tel Aviv University (TAD to DZB). **Author contributions:** NS and DZB designed the experiments. NS performed most of the experiments. DZB wrote the manuscript, incorporating edits and comments from NS.
Competing interests: The authors declare that they have no competing interests.
Data and materials availability: Generated data is available as supplementary files.

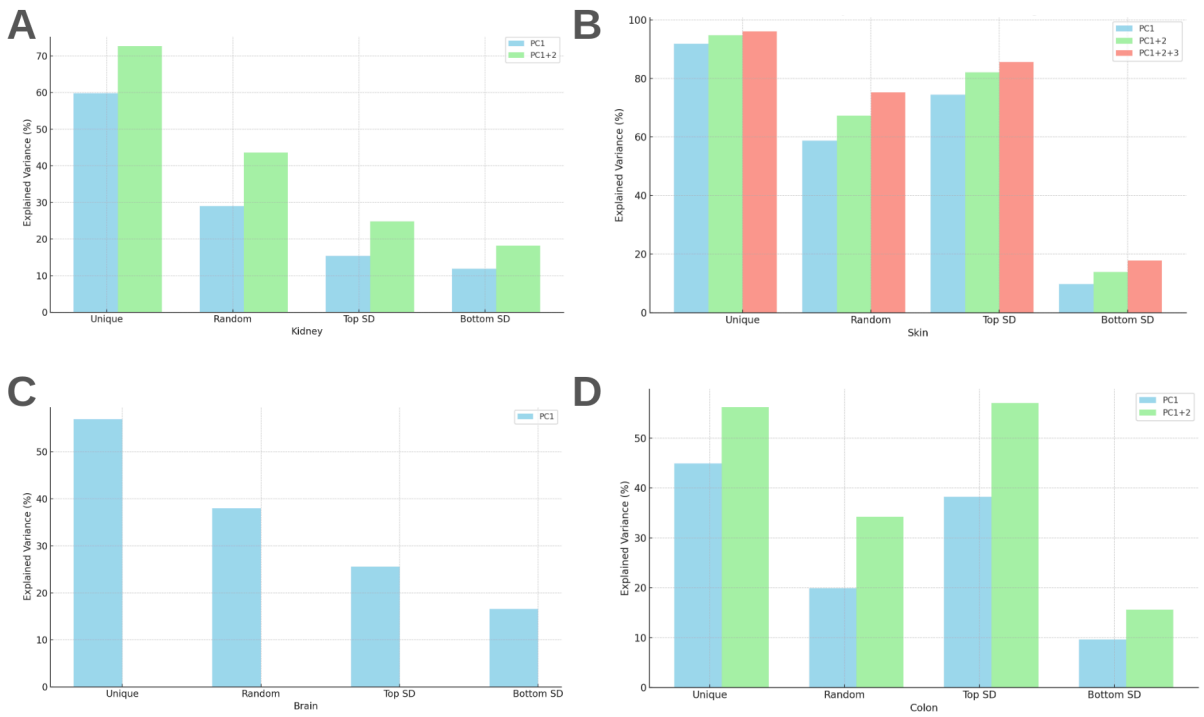
Supplementary Tables

Supplementary Table 1. Tissue unique methylation sites, as identified in 23 of the 30 tissues in the NGDC-CNCB dataset.
Supplementary Table 2. DNA methylation datasets used.
Supplementary Table 3. Tissue-unique methylation data.
Supplementary Table 4. CD4+ unique methylation sites, as identified from the NGDC-CNCB blood dataset.

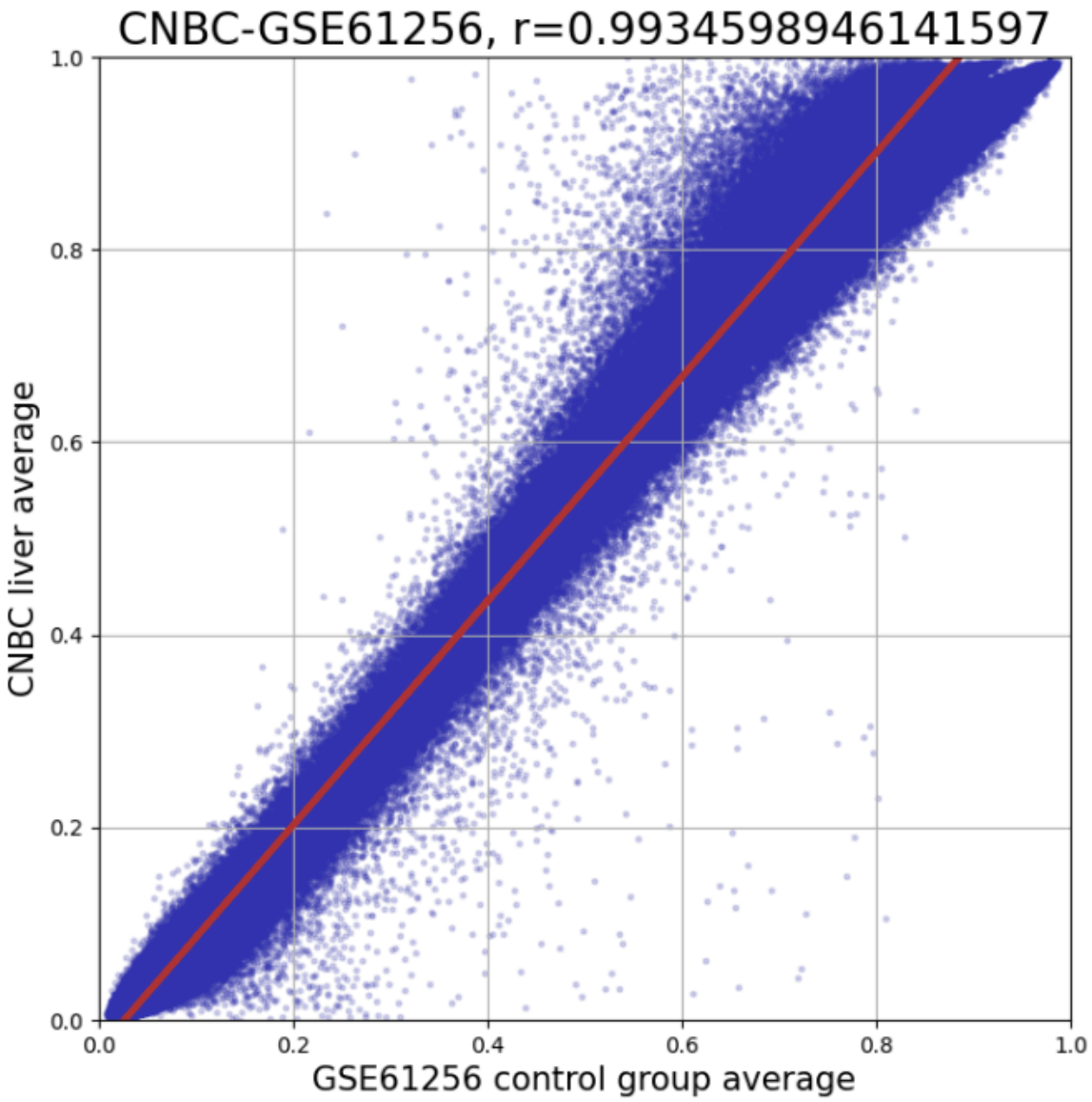
Supplementary figures



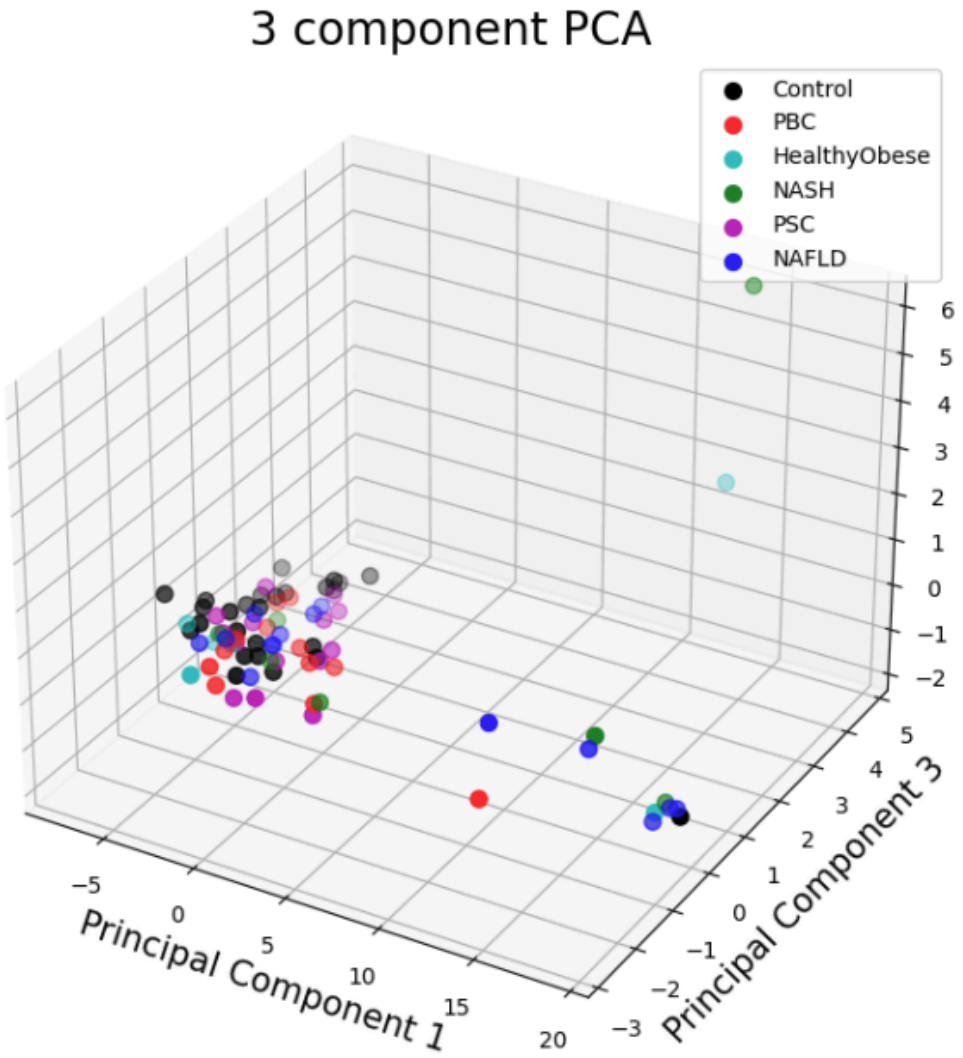
Supplementary Figure 1. TF motif enrichment in different tissues.



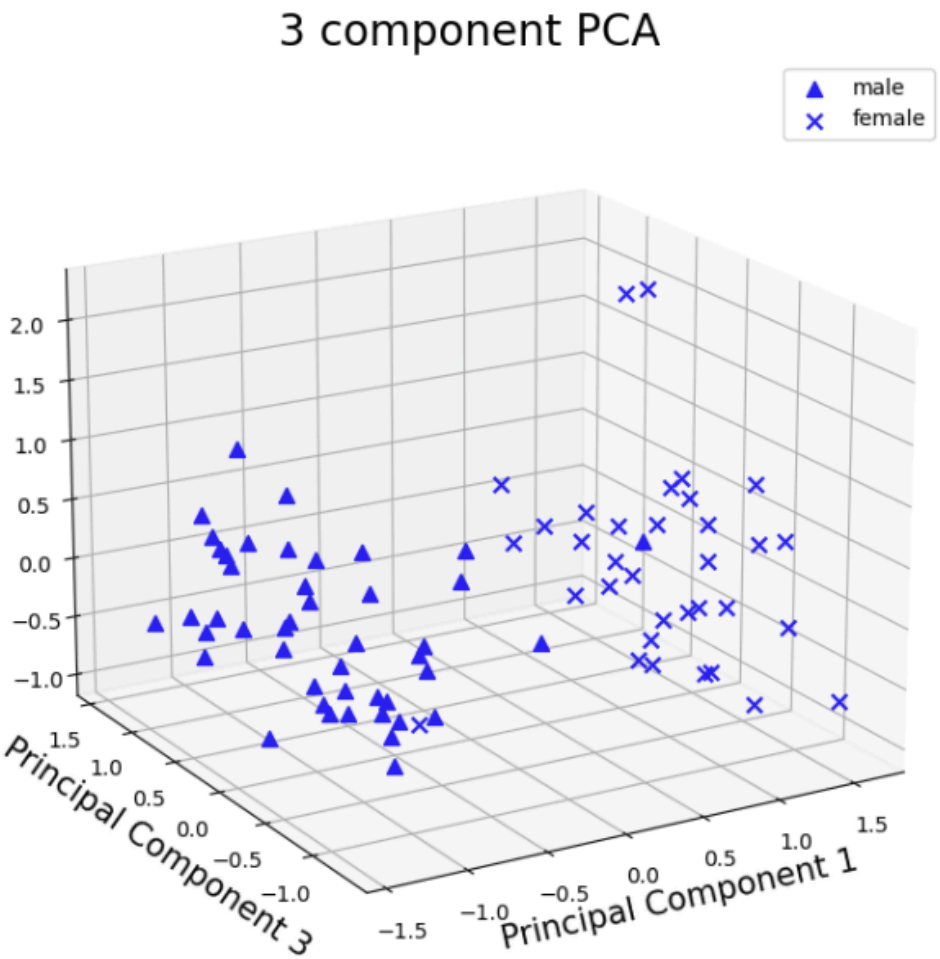
Supplementary Figure 2. Explained variance of principal components. Explained variance of cumulative principal components for kidney (**A**) skin (**B**) brain (**C**) and colon (**D**), using tissue unique, random, top SD and bottom SD sites.



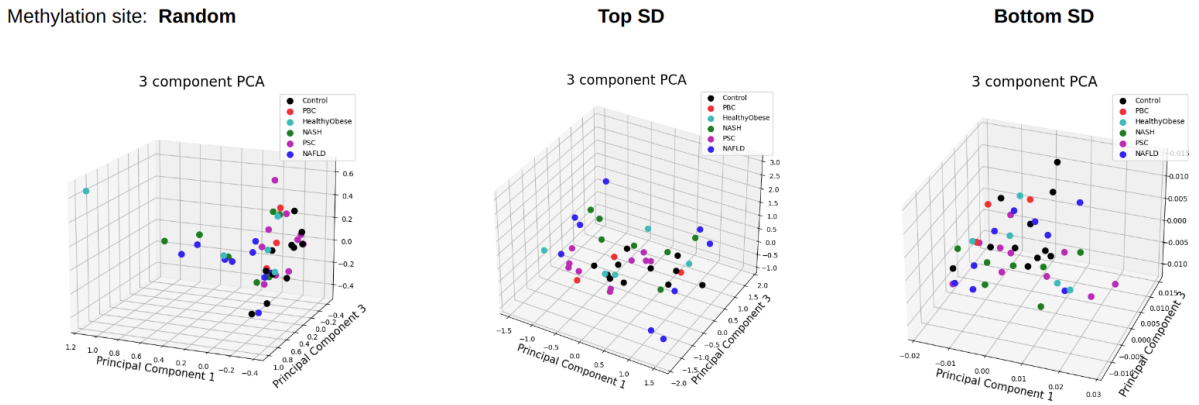
Supplementary Figure 3. Pearson correlation of average liver methylation values in the GSE61256 and the NGDC-CNCB datasets.



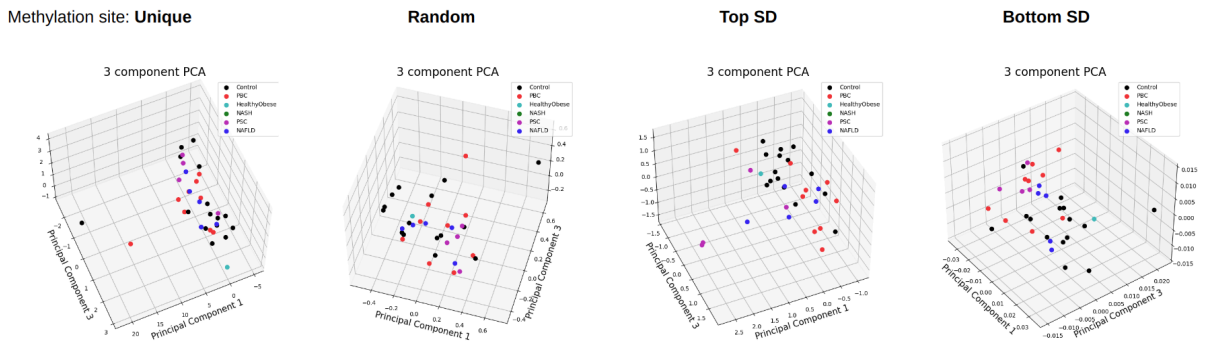
Supplementary Figure 4. PCA analysis of all GSE61256 liver samples, using uniquely methylated sites.



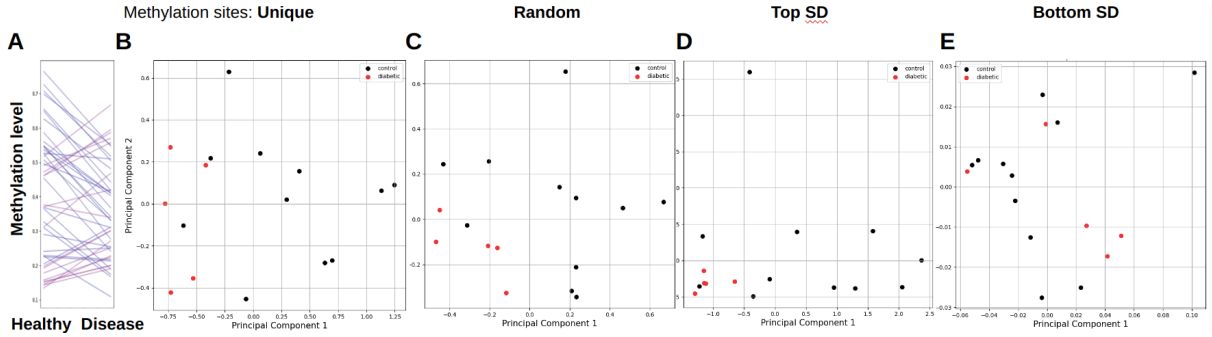
Supplementary Figure 5. PCA analysis of all GSE61256 liver samples using high-variability sites.



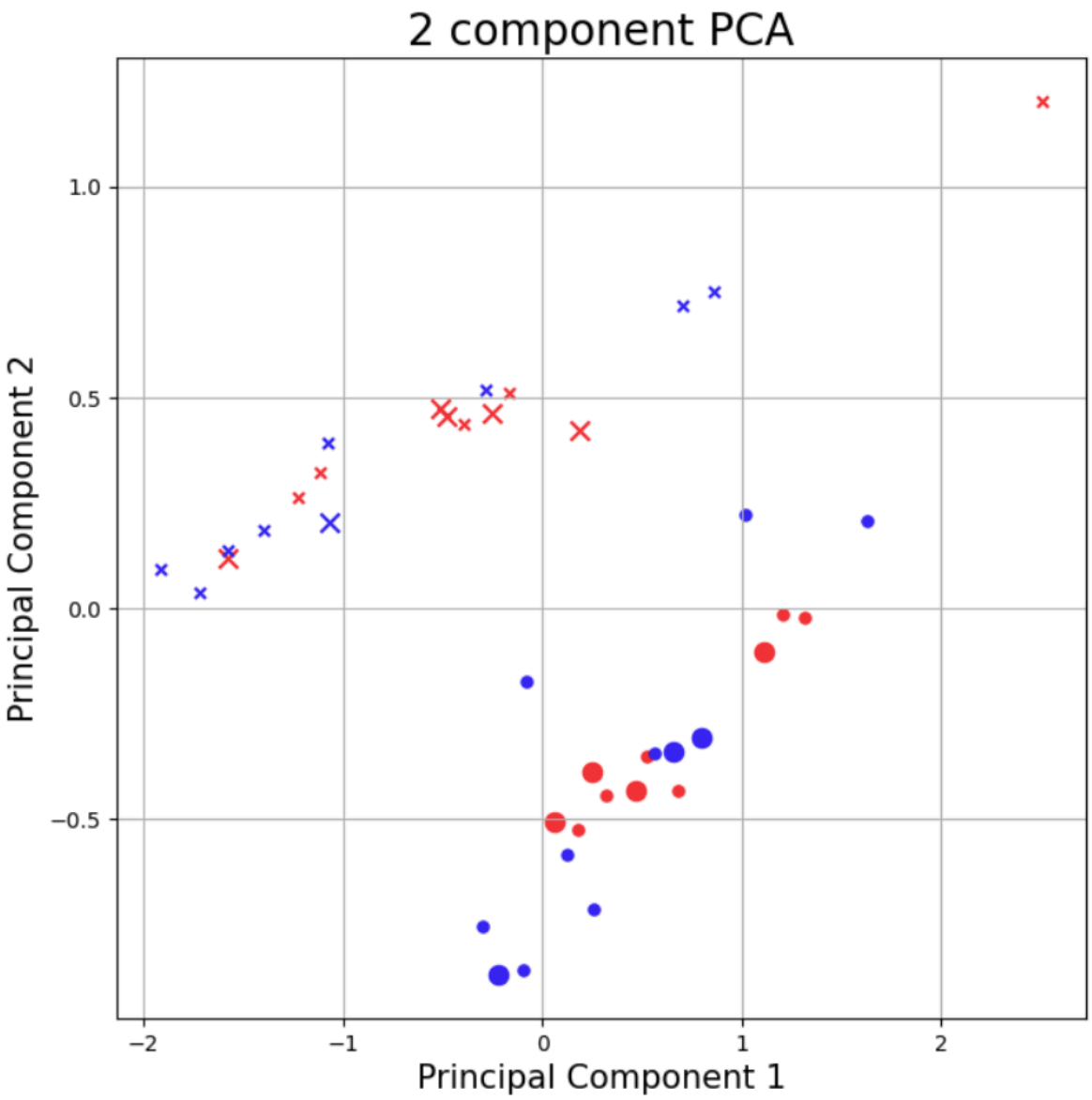
Supplementary Figure 6. PCA analysis of male liver samples from GSE61256, using random, low-variability and high-variability methylated sites.



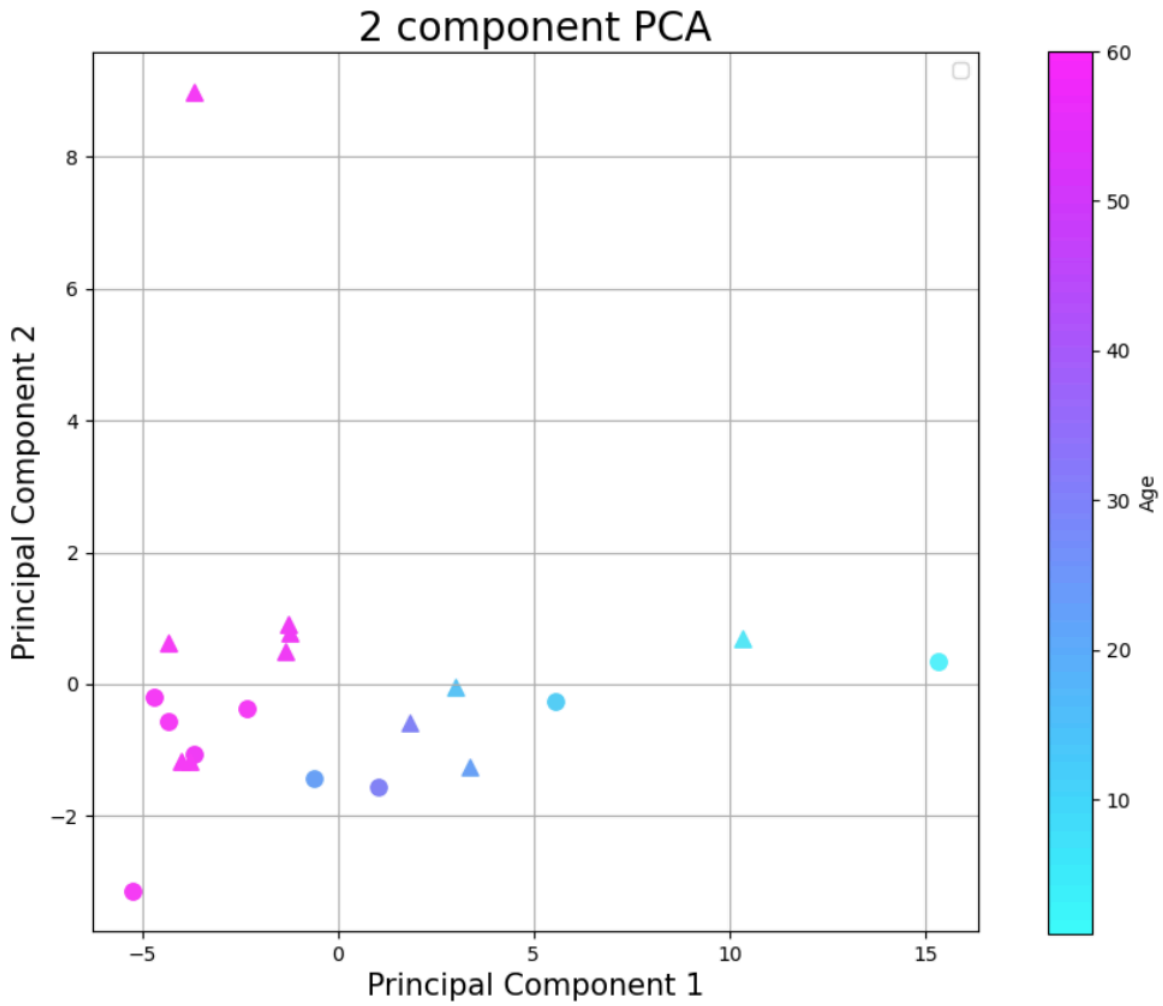
Supplementary Figure 7. PCA analysis of female liver samples from GSE61256, using unique, random, low-variability and high-variability methylated sites.



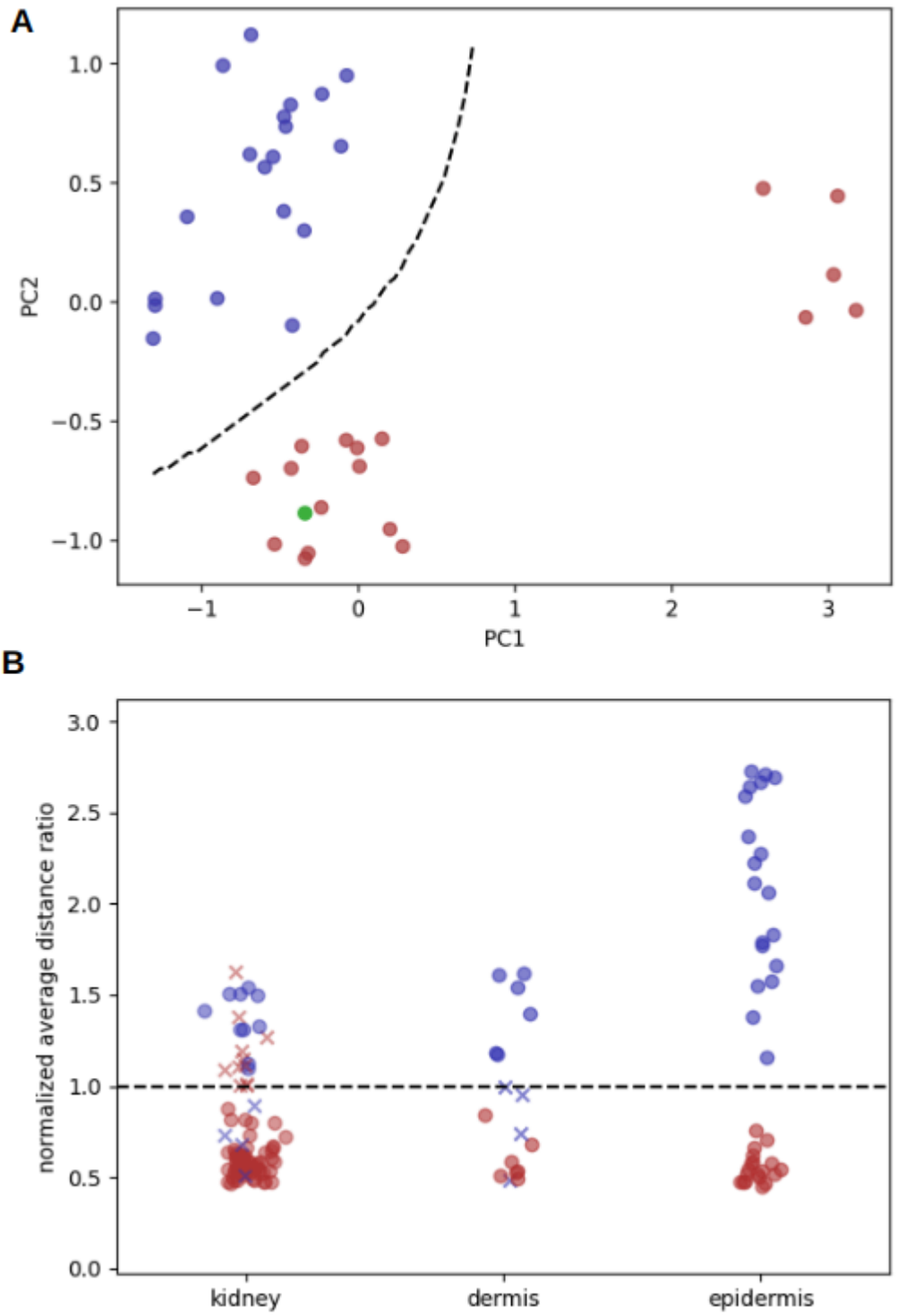
Supplementary Figure 8. Tissue-unique sites in the pancreas of T2D patients. (**A**) Pancreas-unique methylation sites diverge from the mean both for sites that are more (red) and less (blue) methylated. PCA analysis separates healthy (black) from T2D pancreas when using pancreas-unique (**B**) random (**C**) and high variability (**D**) sites, but not low variability sites (**E**).



Supplementary Figure 9. PCA mapping using tissue-unique sites in the adipose of T2D patients. Healthy - blue; diabetic - red. SAT - x; VAT - circle. Size is sex: small - female, large - male.



Supplementary Figure 10. PCA analysis of brain samples from GSE38608, using uniquely methylated sites. Circles represent control samples, while triangles represent autism samples. Colors indicate age.



Supplementary Figure 11. (A) Classification of a single epidermis sample (green) in leave-one-out analysis using 2D PCA. The separating (dashed) line was calculated using the formula $\frac{\sum_{i=1}^{n_1}\sqrt{(pc1-x_{i,ctrl})^2+(pc2-y_{i,ctrl})^2}}{n_1}=\frac{\sum_{i=1}^{n_2}\sqrt{(pc1-x_{i,exp})^2+(pc2-y_{i,exp})^2}}{n_2}$

(B) Kidney filtration rate and skin exposure classification using PCA and average distance ratio in leave-one-out cross validation analysis. Red dots show high eGFR/ sun exposed skin, while blue dots show low eGFR/ sun protected skin. Correct decisions are marked by full circles whereas incorrect decisions are marked by x.