Genome Biology

RESEARCH Open Access

Screening for genes that accelerate the epigenetic aging clock in humans reveals a role for the H3K36 methyltransferase NSD1

Daniel E. Martin-Herranz[1,2*], Erfan Aref-Eshghi[3,4], Marc Jan Bonder[1,5], Thomas M. Stubbs[2], Sanaa Choufani[6], Rosanna Weksberg[6], Oliver Stegle[1,5,7], Bekim Sadikovic[3,4], Wolf Reik[8,9,10*†] and Janet M. Thornton[1*†]

Abstract

Background: Epigenetic clocks are mathematical models that predict the biological age of an individual using DNA methylation data and have emerged in the last few years as the most accurate biomarkers of the aging process. However, little is known about the molecular mechanisms that control the rate of such clocks. Here, we have examined the human epigenetic clock in patients with a variety of developmental disorders, harboring mutations in proteins of the epigenetic machinery.

Results: Using the Horvath epigenetic clock, we perform an unbiased screen for epigenetic age acceleration in the blood of these patients. We demonstrate that loss-of-function mutations in the H3K36 histone methyltransferase NSD1, which cause Sotos syndrome, substantially accelerate epigenetic aging. Furthermore, we show that the normal aging process and Sotos syndrome share methylation changes and the genomic context in which they occur. Finally, we found that the Horvath clock CpG sites are characterized by a higher Shannon methylation entropy when compared with the rest of the genome, which is dramatically decreased in Sotos syndrome patients.

Conclusions: These results suggest that the H3K36 methylation machinery is a key component of the *epigenetic maintenance system* in humans, which controls the rate of epigenetic aging, and this role seems to be conserved in model organisms. Our observations provide novel insights into the mechanisms behind the epigenetic aging clock and we expect will shed light on the different processes that erode the human epigenetic landscape during aging.

Keywords: Aging, Epigenetics, DNA methylation, Epigenetic clock, Biological age, Developmental disorder, Sotos syndrome, H3K36 methylation, NSD1, Methylation entropy

Background

Aging is normally defined as the time-dependent functional decline which increases vulnerability to common diseases and death in most organisms [1]. However, the molecular processes that drive the emergence of age-related diseases are only beginning to be elucidated. With the passage of time, dramatic and complex changes accumulate in the epigenome of cells, from yeast to humans, pinpointing epigenetic alterations as one of the hallmarks of aging [1–4].

Our understanding of the aging process has historically been hampered by the lack of tools to accurately measure it. In recent years, epigenetic clocks have emerged as powerful biomarkers of the aging process across mammals [5, 6], including humans [7–9], mouse [10–14], dogs and wolves [15], and humpback whales [16]. Epigenetic clocks are mathematical models that are trained to predict chronological age using the DNA methylation status of a small number of CpG sites in the genome. The most widely used multi-tissue epigenetic

* Correspondence: dem44@ebi.ac.uk; wolf.reik@babraham.ac.uk; thornton@ebi.ac.uk
†Wolf Reik and Janet M. Thornton jointly supervised this work.
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[8]Epigenetics Programme, The Babraham Institute, Cambridge, UK
Full list of author information is available at the end of the article

clock in humans was developed by Steve Horvath in 2013 [8]. Interestingly, deviations of the epigenetic (biological) age from the expected chronological age (aka epigenetic age acceleration or EAA) have been associated with many conditions in humans, including time-to-death [17, 18], HIV infection [19], Down syndrome [20], obesity [21], Werner syndrome [22], and Huntington's disease [23]. On the contrary, children with multifocal developmental dysfunctions (syndrome X), which seem to evade aging, did not display slower epigenetic aging in a previous study [24]. In mice, the epigenetic clock is slowed down by dwarfism and calorie restriction [11–14, 25] and is accelerated by ovariectomy and high-fat diet [10, 13]. Furthermore, in vitro reprogramming of somatic cells into iPSCs reduces epigenetic age to values close to zero both in humans [8] and mice [11, 14], which opens the door to potential rejuvenation therapies [26, 27].

Epigenetic clocks can be understood as a proxy to quantify the changes of the epigenome with age. However, little is known about the molecular mechanisms that determine the rate of these clocks. Steve Horvath proposed that the multi-tissue epigenetic clock captures the workings of an *epigenetic maintenance system* [8]. Recent GWAS studies have found several genetic variants associated with epigenetic age acceleration in genes such as *TERT* (the catalytic subunit of telomerase) [28], *DHX57* (an ATP-dependent RNA helicase) [29], or *MLST8* (a subunit of both mTORC1 and mTORC2 complexes) [29]. Nevertheless, to our knowledge, no genetic variants in epigenetic modifiers have been found and the molecular nature of this hypothetical system is unknown to this date.

We decided to take a reverse genetics approach and look at the behavior of the epigenetic clock in patients with developmental disorders, many of which harbor mutations in proteins of the epigenetic machinery [30, 31]. We performed an unbiased screen for epigenetic age acceleration and found that Sotos syndrome accelerates epigenetic aging, potentially revealing a role of H3K36 methylation maintenance in the regulation of the rate of the epigenetic clock.

Results

Screening for epigenetic age acceleration is improved when correcting for batch effects

The main goal of this study is to identify genes, mainly components of the epigenetic machinery, that can affect the rate of epigenetic aging in humans (as measured by Horvath's epigenetic clock) [8]. For this purpose, we conducted an unbiased screen for epigenetic age acceleration (EAA) in samples from patients with developmental disorders that we could access and for which genome-wide DNA methylation data was available (Table 1, Additional file 2). Horvath's epigenetic clock, unlike other epigenetic clocks available in the literature, works across the entire human lifespan (even in prenatal samples), and it is therefore well suited for this type of analysis [5, 8, 32]. All the DNA methylation data were generated from the blood using the Illumina HumanMethylation450 array (450K array).

The main step in the screening methodology is to compare the EAA distribution for the samples with a given developmental disorder against a robust control (Fig. 1a). In our case, the control set was obtained from human blood samples in a healthy population of individuals that matched the age range of the developmental disorder samples (Additional file 3). Given that the EAA reflects deviations between the epigenetic (biological) age and the chronological age of a sample, we would expect the EAA distributions of the controls to be centered around zero, which is equivalent to the situation when the median absolute error (MAE) of the model prediction is close to zero (see the "Methods" section). This was not the case for the samples obtained from several control batches (Additional file 1: Figure S1A, S1B), both in the case of EAA models with and without cell composition correction (CCC). It is worth noting that these results were obtained even after applying the internal normalization step against a blood gold standard suggested by Horvath [8]. Therefore, we hypothesized that part of the deviations observed might be caused by technical variance that was affecting epigenetic age predictions in the different batches.

We decided to correct for the potential batch effects by making use of the control probes present on the 450K array, which have been shown to carry information about unwanted variation from a technical source (i.e., technical variance) [33–35]. Performing principal components analysis (PCA) on the raw intensities of the control probes showed that the first two components (PCs) capture the batch structure in both controls (Fig. 1b) and cases (Additional file 1: Figure S1C). Including the first 17 PCs as part of the EAA modeling strategy (see the "Methods" section), which together accounted for 98.06% of the technical variance in controls and cases (Additional file 1: Figure S1D), significantly reduced the median absolute error (MAE) of the predictions in the controls ($\text{MAE}_{\text{without CCC}} = 2.8211$ years, $\text{MAE}_{\text{with CCC}} = 2.7117$ years, mean MAE = 2.7664 years, Fig. 1c). These values are below the original MAE reported by Horvath in his test set (3.6 years) [8].

Finally, deviations from a median EAA close to zero in some of the control batches after batch effect correction (Fig. 1d, Additional file 1: Figure S1E) could be explained by other variables, such as a small batch size or an overrepresentation of young samples (Additional file 1: Figure S1F). The latter is a consequence of the fact that Horvath's model underestimates the epigenetic ages of

Table 1 Overview of the developmental disorders that were included in the screening (total $N = 367$) after quality control (QC) and filtering (see the "Methods" section and Fig. 1a)

Developmental disorder	Gene(s) involved	Gene(s) function	Molecular cause	Number	Age range (years)
Angelman	*UBE3A*	Ubiquitin-protein ligase E3A	Imprinting, mutation	14	1 to 55
Autism spectrum disorder (ASD)	–	–	–	119	1.83 to 35.16
Alpha thalassemia/mental r etardation X-linked syndrome (ATR-X)	*ATRX*	Chromatin remodeling	Mutation	15	0.7 to 27
Claes-Jensen	*KDM5C*	H3K4 demethylase	Mutation	10	2 to 42
Coffin-Lowry	*RPS6KA3*	Serine/threonine kinase	Mutation	10	1.3 to 22.8
Floating-Harbor	*SRCAP*	Chromatin remodeling	Mutation	17	4 to 42
Fragile X syndrome (FXS)	*FMR1*	Translational control	Mutation (CGG expansion)	32	0.08 to 48
Kabuki	*KMT2D*	H3K4 methyltransferase	Mutation	46	0 to 24.1
Noonan	*PTPN11, RAF1, SOS1*	RAS/MAPK signaling	Mutation	15, 11, 14	0.2 to 49
Rett	*MECP2*	Transcriptional repression	Mutation	15	1 to 34
Saethre-Chotzen	*TWIST1*	Transcription factor	Mutation	22	0 to 38
Sotos	*NSD1*	H3K36 methyltransferase	Mutation	20	1.6 to 41
Weaver	*EZH2*	H3K27 methyltransferase	Mutation	7	2.58 to 43

older samples, a phenomenon which has also been observed by other authors [36, 37]. If there is a high number of old samples (generally > 60 years) in the control model, this can lead to a lower model slope, which would incorrectly assign negative EAA to young samples. This highlights the importance of having an age distribution in the control samples that matches that of the cases to be tested for differences in EAA.

Thus, we have shown that correcting for batch effects in the context of the epigenetic clock is important, especially when combining datasets from different sources for meta-analysis purposes. Batch effect correction is essential to remove technical variance that could affect the epigenetic age of the samples and confound biological interpretation.

Sotos syndrome accelerates epigenetic aging

Once we had corrected for potential batch effects in the data, we compared the epigenetic age acceleration (EAA) distributions between each of the developmental disorders studied and our control set. For a given sample, a positive EAA indicates that the epigenetic (biological) age of the sample is higher than the one expected for someone with that chronological age. In other words, it means that the epigenome of that person resembles the epigenome of an older individual. The opposite is true when a negative EAA is found (i.e., the epigenome looks younger than expected).

For the main screen, we selected those control samples with the same age range as the one present when aggregating all the cases (0 to 55 years), since this permits the development of a common control (background) model and to compare the statistical significance of the results across developmental disorders. Only those developmental disorders that satisfied our filtering criteria were considered for the screen (at least 5 samples available for the developmental disorder, with 2 of them presenting a chronological age ≥ 20 years, Fig. 1a, Table 1 and Additional file 2). Given that the blood composition changes with age (changes in the different cell type proportions, which can affect bulk DNA methylation measurements), we used models with and without cell composition correction (CCC), correcting for batch effects in both of them (see the "Methods" section). It is important to mention that $\mathrm{EAA}_{\mathrm{with\ CCC}}$ is conceptually similar to the previously reported measure of "intrinsic EAA" (IEAA) [18, 38].

The results from the screen are portrayed in Fig. 2a. Most syndromes do not show evidence of accelerated epigenetic aging, but Sotos syndrome presents a clear positive EAA (median $\mathrm{EAA}_{\mathrm{with\ CCC}} = +7.64$ years, median $\mathrm{EAA}_{\mathrm{without\ CCC}} = +7.16$ years), with p values considerably below the significance level of 0.01 after Bonferroni correction (p $\mathrm{value}_{\mathrm{corrected,\ with\ CCC}} = 3.40 \times 10^{-9}$, p $\mathrm{value}_{\mathrm{corrected,\ without\ CCC}} = 2.61 \times 10^{-7}$). Additionally, Rett syndrome (median $\mathrm{EAA}_{\mathrm{with\ CCC}} = +2.68$ years, median $\mathrm{EAA}_{\mathrm{without\ CCC}} = +2.46$ years, p $\mathrm{value}_{\mathrm{corrected,\ with\ CCC}} = 0.0069$, p $\mathrm{value}_{\mathrm{corrected,\ without\ CCC}} = 0.0251$) and Kabuki syndrome (median $\mathrm{EAA}_{\mathrm{with\ CCC}} = -1.78$ years, median $\mathrm{EAA}_{\mathrm{without\ CCC}} = -2.25$ years, p $\mathrm{value}_{\mathrm{corrected,\ with\ CCC}} = 0.0011$, p $\mathrm{value}_{\mathrm{corrected,\ without\ CCC}} = 0.0035$) reach significance, with a positive and negative EAA,

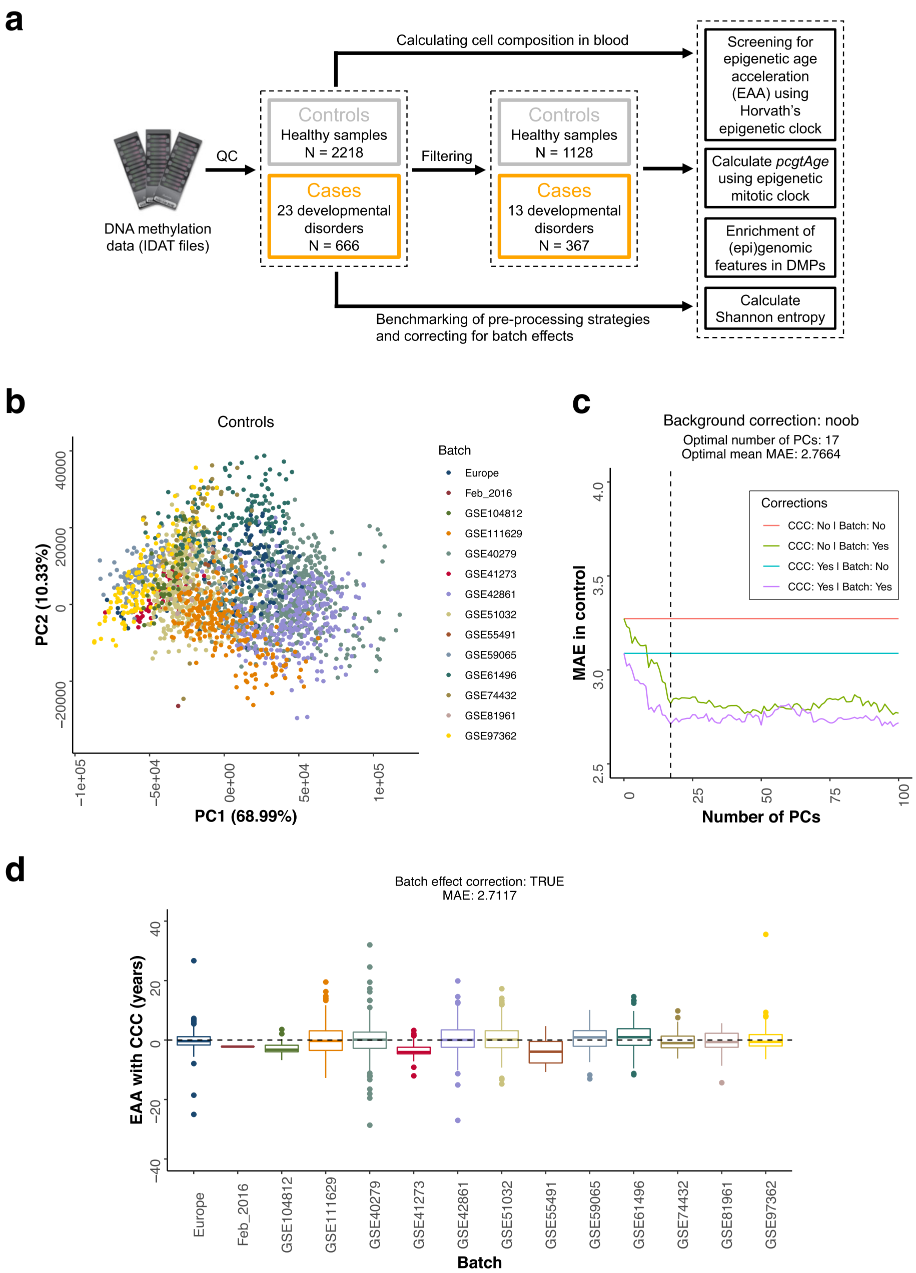


Fig. 1 Screening for epigenetic age acceleration (EAA) is improved when correcting for batch effects. **a** Flow diagram that portrays an overview of the different analyses that are carried out in the raw DNA methylation data (IDAT files) from human blood for cases (developmental disorders samples) and controls (healthy samples). The control samples are filtered to match the age range of the cases (0–55 years). The cases are filtered based on the number of "adult" samples available (for each disorder, at least 5 samples, with 2 of them with an age ≥ 20 years). More details can be found in the "Methods" section. QC, quality control; DMPs, differentially methylated positions. **b** Scatterplot showing the values of the first two principal components (PCs) for the control samples after performing PCA on the control probes of the 450K arrays. Each point corresponds to a different control sample, and the colors represent the different batches. The different batches cluster together in the PCA space, showing that the control probes indeed capture technical variation. Please note that all the PCA calculations were done with more samples from cases and controls than those that were included in the final screening since it was performed before the filtering step (see the "Methods" section and Fig. 1a). **c** Plot showing how the median absolute error (MAE) of the prediction in the control samples, that should tend to zero, is reduced when the PCs capturing the technical variation are included as part of the modeling strategy (see the "Methods" section). The dashed line represents the optimal number of PCs (17) that was finally used. The optimal mean MAE is calculated as the average MAE between the green and purple lines. CCC, cell composition correction. **d** Distribution of the EAA with cell composition correction (CCC) for the different control batches, after applying batch effect correction

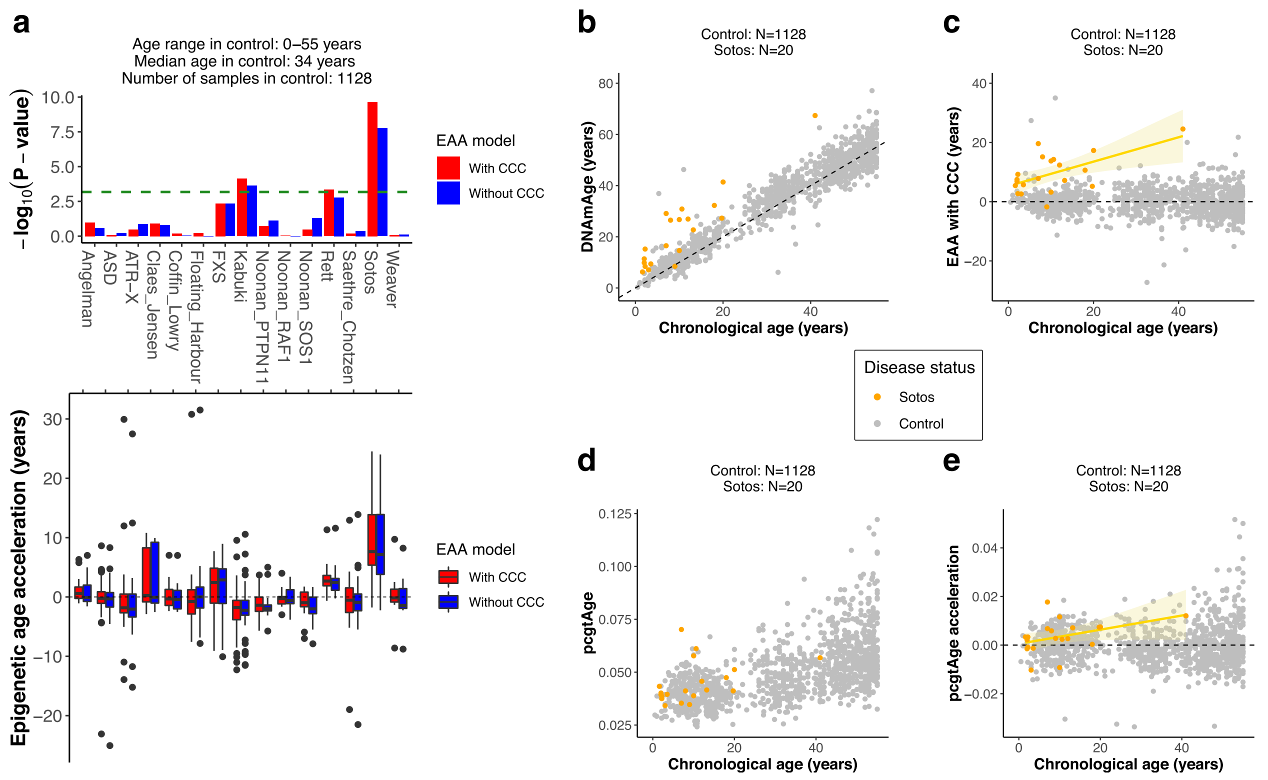


Fig. 2 Sotos syndrome accelerates epigenetic aging. **a** Screening for epigenetic age acceleration (EAA) in developmental disorders. The upper panel shows the *p* values derived from comparing the EAA distributions for the samples in a given developmental disorder and the control (two-sided Wilcoxon's test). The dashed green line displays the significance level of $\alpha = 0.01$ after Bonferroni correction. The bars above the green line reach statistical significance. The lower panel displays the actual EAA distributions, which allows assessing the direction of the EAA (positive or negative). In red: EAA model with cell composition correction (CCC). In blue: EAA model without CCC. ASD, autism spectrum disorder; ATR-X, alpha thalassemia/ mental retardation X-linked syndrome; FXS, fragile X syndrome. **b** Scatterplot showing the relation between epigenetic age (*DNAmAge*) according to Horvath's model [8] and chronological age of the samples for Sotos (orange) and control (gray). Each sample is represented by one point. The black dashed line represents the diagonal to aid visualization. **c** Scatterplot showing the relation between the epigenetic age acceleration (EAA) and chronological age of the samples for Sotos (orange) and control (gray). Each sample is represented by one point. The yellow line represents the linear model EAA ~ Age, with the standard error shown in the light yellow shade. **d** Scatterplot showing the relation between the score for the epigenetic mitotic clock (*pcgtAge*) [39] and chronological age of the samples for Sotos (orange) and control (gray). Each sample is represented by one point. A higher value of *pcgtAge* is associated with a higher number of cell divisions in the tissue. **e** Scatterplot showing the relation between the epigenetic mitotic clock (*pcgtAge*) acceleration and chronological age of the samples for Sotos (orange) and control (gray). Each sample is represented by one point. The yellow line represents the linear model $pcgtAge_{\mathrm{acceleration}}$ ~ Age, with the standard error shown in the light yellow shade

respectively. Finally, fragile X syndrome (FXS) shows a positive EAA trend (median $\mathrm{EAA}_{\mathrm{with\ CCC}} = +2.44$ years, median $\mathrm{EAA}_{\mathrm{without\ CCC}} = +2.88$ years) that does not reach significance in our screen (p $\mathrm{value}_{\mathrm{corrected,\ with\ CCC}} = 0.0680$, p $\mathrm{value}_{\mathrm{corrected,\ without\ CCC}} = 0.0693$).

Next, we tested the effect of changing the median age used to build the healthy control model (i.e., the median age of the controls) on the screening results (Additional file 1: Figure S2A). Sotos syndrome is robust to these changes, whilst Rett, Kabuki, and FXS are much more sensitive to the control model used. This again highlights the importance of choosing an appropriate age-matched control when testing for epigenetic age acceleration, given that Horvath's epigenetic clock underestimates epigenetic age for advanced chronological ages [36, 37].

Moreover, all but one of the Sotos syndrome patients (19/20 = 95%) show a consistent deviation in EAA (with CCC) in the same direction (Fig. 2b, c), which is not the case for the rest of the disorders, with the exception of Rett syndrome (Additional file 1: Figure S2B). Even though the data suggest that there are already some methylomic changes at birth, the EAA seems to increase with age in the case of Sotos patients (Fig. 2c; *p* values for the slope coefficient of the EAA ~ Age linear regression: p $\mathrm{value}_{\mathrm{with\ CCC}} = 0.00569$, p $\mathrm{value}_{\mathrm{without\ CCC}} = 0.00514$). This could imply that at least some of the changes that normally affect the epigenome with age are happening at a faster rate in Sotos syndrome patients during their lifespan (as opposed to the idea that the Sotos epigenetic changes are only acquired during prenatal development and remain constant afterwards).

Nevertheless, this increase in EAA with chronological age is highly influenced by a single patient with a chronological age of 41 years (i.e., if this patient is removed, the p values for the slope coefficient are p value$_{\text{with CCC}} = 0.1785$ and p value$_{\text{without CCC}} = 0.1087$ respectively). Therefore, more data of older Sotos patients are required to be certain about the dynamics of these methylomic changes.

In order to further validate the epigenetic age acceleration observed in Sotos patients, we calculated their epigenetic age according to other widely used epigenetic clocks: Hannum's clock [9], Lin's clock [40], and the skin-blood clock [41]. These analyses confirmed that Sotos patients clearly present accelerated epigenetic aging when compared with healthy individuals (with the exception of the EAA$_{\text{without CCC}}$ in the skin-blood clock, which showed the same trend but did not reach significance; Additional file 1: Figure S2C-E).

Finally, we investigated whether Sotos syndrome leads to a higher rate of (stem) cell division in the blood when compared with our healthy population. We used a reported epigenetic mitotic clock (*pcgtAge*) that makes use of the fact that some CpGs in promoters that are bound by Polycomb group proteins become hypermethylated with age. This hypermethylation correlates with the number of cell divisions in the tissue and is also associated with an increase in cancer risk [39]. We found a trend suggesting that the epigenetic mitotic clock might be accelerated in Sotos patients (p value = 0.0112, Fig. 2d, e), which could explain the higher cancer predisposition reported in these patients and might relate to their overgrowth [42]. Again, this trend could be influenced by the 41-year-old Sotos patient (after removing this patient: p value = 0.0245), and more data of older Sotos patients is required to confirm this observation.

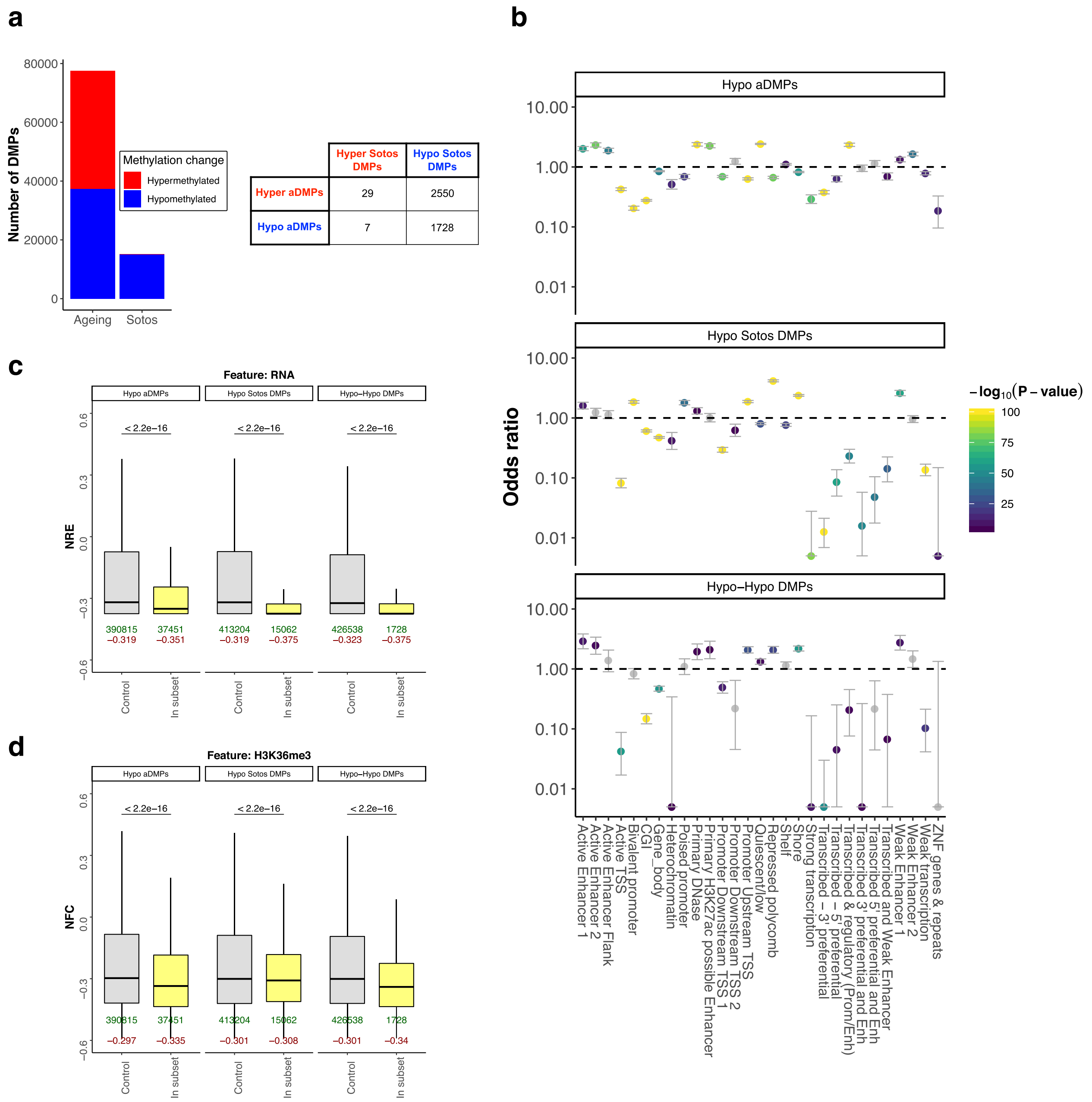
Consequently, we report that individuals with Sotos syndrome present an accelerated epigenetic age, which makes their epigenome look, on average, more than 7 years older than expected. These changes could be the consequence of a higher ticking rate of the epigenetic clock (or at least part of its machinery), with epigenetic age acceleration potentially increasing during lifespan: the youngest Sotos patient (1.6 years) has an EAA$_{\text{with CCC}}$ = 5.43 years and the oldest (41 years) has an EAA$_{\text{with CCC}}$ = 24.53 years. Additionally, Rett syndrome, Kabuki syndrome, and fragile X syndrome could also have their epigenetic ages affected, but more evidence is required to be certain about this conclusion.

Physiological aging and Sotos syndrome share methylation changes and the genomic context in which they occur

Sotos syndrome is caused by loss-of-function heterozygous mutations in the *NSD1* gene, a histone H3K36 methyltransferase [43, 44]. These mutations lead to a specific DNA methylation signature in Sotos patients, potentially due to the crosstalk between the histone and DNA methylation machinery [44]. In order to gain a more detailed picture of the reported epigenetic age acceleration, we decided to compare the genome-wide (or at least array-wide) changes observed in the methylome during aging with those observed in Sotos syndrome. For this purpose, we identified differentially methylated positions (DMPs) for both conditions (see the "Methods" section). Aging DMPs (aDMPs), were composed almost equally of CpG sites that gain methylation with age (i.e., become hypermethylated, 51.69%) and CpG sites that lose methylation with age (i.e., become hypomethylated, 48.31%, barplot in Fig. 3a), a picture that resembles previous studies [45]. On the contrary, DMPs in Sotos were dominated by CpGs that decrease their methylation level in individuals with the syndrome (i.e., hypomethylated, 99.27%, barplot in Fig. 3a), consistent with previous reports [44].

Then, we compared the intersections between the hypermethylated and hypomethylated DMPs in aging and Sotos. Most of the DMPs were specific for aging or Sotos (i.e., they did not overlap), but a subset of them was shared (table in Fig. 3a). Interestingly, there were 1728 DMPs that became hypomethylated both during aging and in Sotos (Hypo-Hypo DMPs). This subset of DMPs is of special interest because it could be used to understand in more depth some of the mechanisms that drive hypomethylation during physiological aging. Thus, we tested whether the different subsets of DMPs are found in specific genomic contexts (Additional file 1: Figure S3A,B). DMPs that are hypomethylated during aging and in Sotos were both enriched (odds ratio > 1) in enhancer categories (such as "active enhancer 1" or "weak enhancer 1", see the chromatin state model used, from the K562 cell line, in the "Methods" section) and depleted (odds ratio < 1) for active transcription categories (such as "active TSS" or "strong transcription"), which was also observed in the "Hypo-Hypo DMPs" subset (Fig. 3b). Interestingly, age-related hypomethylation in enhancers seems to be a characteristic of both humans [46, 47] and mice [25]. Furthermore, both de novo DNA methyltransferases (DNMT3A and DNMT3B) have been shown to bind in an H3K36me3-dependent manner to active enhancers [48], consistent with our results.

When looking at the levels of total RNA expression (depleted for rRNA) in the blood, we confirmed a significant reduction in the RNA levels around these hypomethylated DMPs when compared with the control sets (Fig. 3c, see the "Methods" section for more details on how the control sets were defined). Interestingly, hypomethylated DMPs in both aging and Sotos were depleted from the gene bodies (Fig. 3b) and were located



	Hyper Sotos DMPs	Hypo Sotos DMPs
Hyper aDMPs	29	2550
Hypo aDMPs	7	1728

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Comparison between the DNA methylation changes during physiological aging and in Sotos. **a** Left: barplot showing the total number of differentially methylated positions (DMPs) found during physiological aging and in Sotos syndrome. CpG sites that increase their methylation levels with age in our healthy population or those that are elevated in Sotos patients (when compared with a control) are displayed in red. Conversely, those CpG sites that decrease their methylation levels are displayed in blue. Right: a table that represents the intersection between the aging (aDMPs) and the Sotos DMPs. The subset resulting from the intersection between the hypomethylated DMPs in aging and Sotos is called the "Hypo-Hypo DMPs" subset ($N = 1728$). **b** Enrichment for the categorical (epi) genomic features considered when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in aging and Sotos against a control (see the "Methods" section). The *y*-axis represents the odds ratio (OR), the error bars show the 95% confidence interval for the OR estimate and the color of the points codes for -$\log_{10}$(*p* value) obtained after testing for enrichment using Fisher's exact test. An OR > 1 shows that the given feature is enriched in the subset of DMPs considered, whilst an OR < 1 shows that it is found less than expected. In gray: features that did not reach significance using a significance level of $\alpha = 0.01$ after Bonferroni correction. **c** Boxplots showing the distributions of the "normalised RNA expression" (NRE) when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in aging and Sotos against a control (see the "Methods" section). NRE represents normalized mean transcript abundance in a window of ± 200 bp from the CpG site coordinate (DMP) being considered (see the "Methods" section). The *p* values (two-sided Wilcoxon's test, before multiple testing correction) are shown above the boxplots. The number of DMPs belonging to each subset (in green) and the median value of the feature score (in dark red) are shown below the boxplots. **d** Same as **c**, but showing the "normalised fold change" (NFC) for the H3K36me3 histone modification (representing normalized mean ChIP-seq fold change for H3K36me3 in a window of ± 200 bp from the DMP being considered, see the "Methods" section)

in areas with lower levels of H3K36me3 when compared with the control sets (Fig. 3d, see Additional file 1: Figure S3B for a comprehensive comparison of all the DMPs subsets). Moreover, hypomethylated aDMPs and hypomethylated Sotos DMPs were both generally enriched or depleted for the same histone marks in the blood (Additional file 1: Figure S3B), which adds weight to the hypothesis that they share the same genomic context and could become hypomethylated through similar molecular mechanisms.

Intriguingly, we also identified a subset of DMPs (2550) that were hypermethylated during aging and hypomethylated in Sotos (Fig. 3a). These "Hyper-Hypo DMPs" seem to be enriched for categories such as "bivalent promoter" and 'repressed polycomb' (Additional file 1: Figure S3A), which are normally associated with developmental genes [49, 50]. These categories are also a defining characteristic of the hypermethylated aDMPs, highlighting that even though the direction of the DNA methylation changes is different in some aging and Sotos DMPs, the genomic context in which they happen is shared.

Finally, we looked at the DNA methylation patterns in the 353 Horvath's epigenetic clock CpG sites for the Sotos samples. For each clock CpG site, we modeled the changes of DNA methylation during the lifespan in the healthy control individuals and then calculated the deviations from these patterns for the Sotos samples (Additional file 1: Figure S3C, see the "Methods" section). As expected, the landscape of clock CpG sites is dominated by hypomethylation in the Sotos samples, although only a small fraction of the clock CpG sites seem to be significantly affected (Additional file 1: Figure S3D, Additional file 6). Overall, we confirmed the trends reported for the genome-wide analysis (Additional file 1: Figure S3E-G). However, given the much smaller number of CpG sites to consider in this analysis, very few comparisons reached significance.

We have demonstrated that the aging process and Sotos syndrome share a subset of hypomethylated CpG sites that are characterized by an enrichment in enhancer features and a depletion of active transcription activity. This highlights the usefulness of developmental disorders as a model to study the mechanisms that may drive the changes in the methylome with age, since they permit stratification of the aging DMPs into different functional categories that are associated with alterations in the function of specific genes and hence specific molecular components of the epigenetic aging clock.

Sotos syndrome is associated with a decrease of methylation Shannon entropy in the epigenetic clock CpG sites

Shannon entropy can be used in the context of DNA methylation analysis to estimate the information content stored in a given set of CpG sites. Shannon entropy is minimized when the methylation levels of all the CpG sites are either 0% or 100% and maximized when all of them are 50% (see the "Methods" section). Previous reports have shown that the Shannon entropy associated with the methylome increases with age, which implies that the epigenome loses information content [9, 12, 46]. We confirmed this genome-wide effect (i.e., considering all the CpG sites that passed our pre-processing pipeline) in our healthy samples, where we observed a positive Spearman correlation coefficient between chronological age and genome-wide Shannon entropy of 0.3984 (p value = 3.21×10^{-44}). This result was robust when removing outlier batches (Additional file 1: Figure S4C). Next, we tested whether Sotos patients present genome-wide Shannon entropy acceleration, i.e., deviations from the expected genome-wide Shannon entropy for their age (see the "Methods" section). Despite detailed analysis, we did not find evidence that this was the case when looking genome-wide (p value = 0.71, Fig. 4a, b; Additional file 1: Figure S4A). This conclusion held

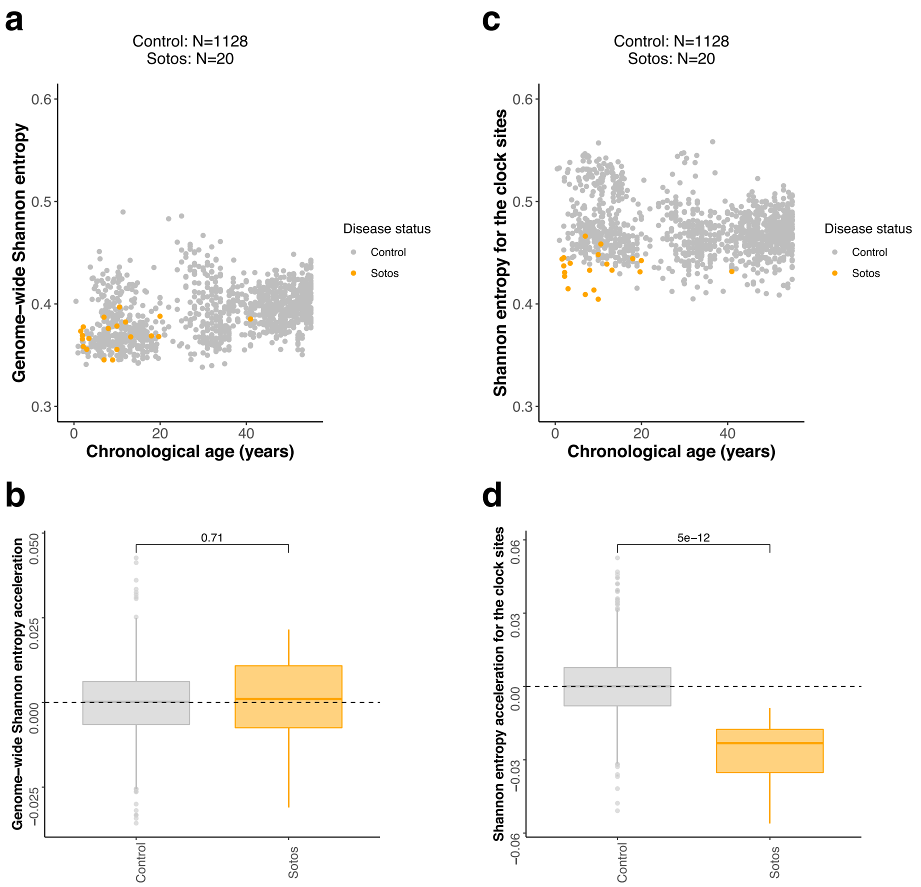


Fig. 4 Analysis of methylation Shannon entropy during physiological aging and in Sotos syndrome. **a** Scatterplot showing the relation between genome-wide Shannon entropy (i.e., calculated using the methylation levels of all the CpG sites in the array) and chronological age of the samples for Sotos (orange) and healthy controls (gray). Each sample is represented by one point. **b** Boxplots showing the distributions of genome-wide Shannon entropy acceleration (i.e., deviations from the expected genome-wide Shannon entropy for their age) for the control and Sotos samples. The *p* value displayed on top of the boxplots was derived from a two-sided Wilcoxon's test. **c** Same as **a**., but using the Shannon entropy calculated only for the 353 CpG sites in the Horvath epigenetic clock. **d** Same as **b**, but using the Shannon entropy calculated only for the 353 CpG sites in the Horvath epigenetic clock

when the comparison was performed inside the batch that contained the Sotos samples (GSE74432), therefore providing evidence that it is not confounded by batch effect (p value = 0.73, Additional file 1: Figure S4E).

When we considered only the 353 clock CpG sites for the entropy calculations, the picture was different. Shannon entropy for the 353 clock sites slightly decreased with age in the controls when we included all the batches, showing the opposite direction when compared with the genome-wide entropy (Spearman correlation coefficient = – 0.1223, p value = 3.8166×10^{-5}, Fig. 4c). However, when we removed the "Europe" batch (which was an outlier even after pre-processing, Additional file 1: Figure S4D), this trend was reversed and we observed a weak increase of clock Shannon entropy with age (Spearman correlation coefficient = 0.1048, p value = 8.6245×10^{-5}). This shows that Shannon entropy calculations are very sensitive to batch effects, especially when considering a small number of CpG sites, and the results must be interpreted carefully.

Interestingly, the mean Shannon entropy across all the control samples was higher in the epigenetic clock sites (mean = 0.4726, Fig. 4c) with respect to the genome-wide entropy (mean = 0.3913, Fig. 4a). Sotos syndrome patients displayed a lower clock Shannon entropy when compared with the control (p value = 5.0449×10^{-12}, Fig. 4d, Additional file 1: Figure S4B), which is probably driven by the hypomethylation of the clock CpG sites. Importantly, this conclusion held when the comparison was performed inside the batch that contained the Sotos samples (GSE74432), again providing evidence that it is not confounded by batch effect (p value = 7.3757×10^{-11}, Additional file 1: Figure S4F). Furthermore, this highlights that the Horvath clock sites could have slightly different characteristics in terms of the methylation entropy associated with them when compared with the genome as a whole, something that to our knowledge has not been reported before.

Discussion

The epigenetic aging clock has emerged as the most accurate biomarker of the aging process, and it seems to be a conserved property in mammalian genomes [5, 6]. However, we do not know yet whether the age-related DNA methylation changes measured are functional at all or whether they are related to some fundamental process of the biology of aging. Developmental disorders in humans represent an interesting framework to look at the biological effects of mutations in genes that are fundamental for the integrity of the epigenetic landscape and other core processes, such as growth or neurodevelopment [30, 31]. Furthermore, according to the *epigenetic clock theory of aging*, epigenetic clocks provide a continuous readout that connects purposeful processes in development with adverse effects in later life [5]. Therefore, using a reverse genetics approach, we aimed to identify the genes that disrupt the aspects of the behavior of the epigenetic aging clock in humans.

Most of the studies have looked at the epigenetic aging clock using Horvath's model [8], which has a ready-to-use online calculator for epigenetic age [51]. This has clearly simplified the computational process and helped a lot of research groups to test the behavior of the epigenetic clock in their system of interest. However, this has also led to the treatment of the epigenetic clock as a "black-box", without a critical assessment of the statistical methodology behind it. Therefore, we decided to benchmark the main steps involved when estimating epigenetic age acceleration (pre-processing of the raw data from methylation arrays and cell composition deconvolution algorithms), to quantify the effects of technical variation on the epigenetic clock predictions and to assess the impact of the control age distribution on the epigenetic age acceleration calculations. Previous attempts to account for technical variation have used the first 5 principal components (PCs) estimated directly from the DNA methylation data [23]. However, this approach potentially removes meaningful biological variation. For the first time, we have shown that it is possible to use the control probes from the 450K array to readily correct for batch effects in the context of the epigenetic clock, which reduces the error associated with the predictions and decreases the likelihood of reporting a false positive. Furthermore, we have confirmed the suspicion that Horvath's model underestimates epigenetic age for older ages [36, 37] and assessed the impact of this bias in the screen for epigenetic age acceleration.

The results from our screen strongly suggest that Sotos syndrome accelerates epigenetic aging, and this effect was confirmed using other epigenetic clocks. Sotos syndrome is caused by loss-of-function mutations in the *NSD1* gene [43, 44], which encodes a histone H3 lysine 36 (H3K36) methyltransferase. This leads to a phenotype which can include prenatal and postnatal overgrowth, facial gestalt, advanced bone age, developmental delay, higher cancer predisposition, and, in some cases, heart defects [42]. Remarkably, many of these characteristics could be interpreted as aging-like, identifying Sotos syndrome as a potential human model of accelerated physiological aging.

NSD1 catalyzes the addition of either monomethyl (H3K36me) or dimethyl groups (H3K36me2) and indirectly regulates the levels of trimethylation (H3K36me3) by altering the availability of the monomethyl and dimethyl substrates for the trimethylation enzymes (SETD2 in humans, whose mutations cause a "Sotos-like" overgrowth syndrome) [52, 53]. H3K36 methylation has a complex role in the regulation of transcription [52] and has been shown to regulate nutrient stress response in yeast [54]. Moreover, experiments in model organisms (yeast and worm) have demonstrated that mutations in H3K36 methyltranferases decrease lifespan, and remarkably, mutations in H3K36 demethylases increase it [55–57].

In humans, DNA methylation patterns are established and maintained by three conserved enzymes: the maintenance DNA methyltransferase DNMT1 and the de novo DNA methyltransferases DNMT3A and DNMT3B [58]. Both DNMT3A and DNMT3B contain PWWP domains that can read the H3K36me3 histone mark [59, 60]. Therefore, the H3K36 methylation landscape can influence DNA methylation levels in specific genomic regions through the recruitment of the de novo DNA methyltransferases. Mutations in the PWWP domain of DNMT3A impair its binding to H3K36me2 and H3K36me3 and cause an undergrowth disorder in humans (microcephalic dwarfism) [61]. This redirects DNMT3A, which is normally targeted to H3K36me2 and H3K36me3 throughout the genome, to DNA methylation valleys (DMVs, aka DNA methylation canyons), which become hypermethylated [61], a phenomenon that also seems to happen during physiological aging in humans [46, 62, 63] and mice [25]. DMVs are hypomethylated domains conserved across cell types and species, often associated with Polycomb-regulated developmental genes and marked by bivalent chromatin (with H3K27me3 and H3K4me3) [64–67]. Therefore, we suggest a model (Fig. 5) where the reduction in the levels of H3K36me2 and/or H3K36me3, caused by a proposed decrease in H3K36 methylation maintenance during aging or NSD1 function in Sotos syndrome, could lead to hypomethylation in many genomic regions (because DNMT3A is recruited less efficiently) and hypermethylation in DMVs (because of the higher availability of DNMT3A). Indeed, we observe enrichment for

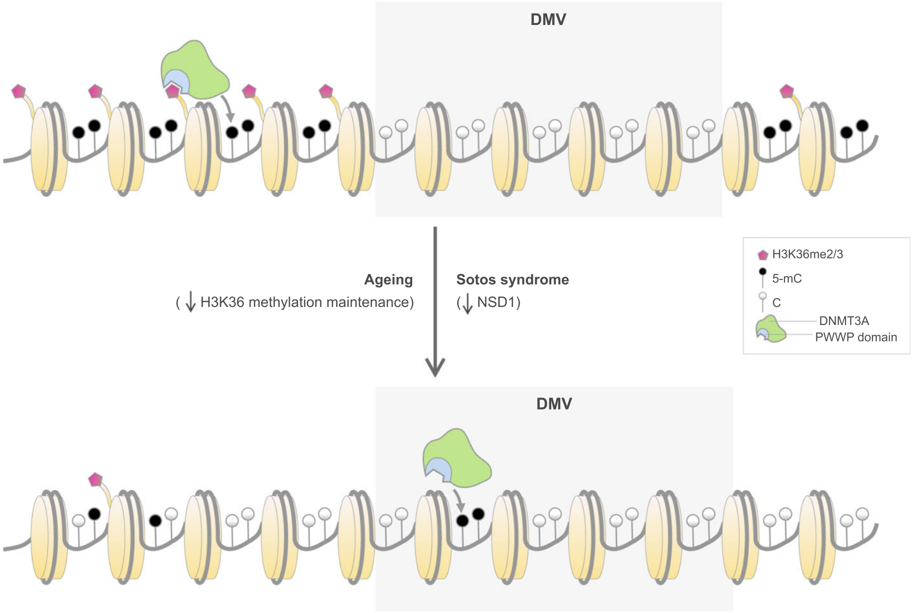


Fig. 5 Proposed model that highlights the role of H3K36 methylation maintenance on epigenetic aging. The H3K36me2/3 mark allows recruiting de novo DNA methyltransferases DNMT3A (in green) and DNMT3B (not shown) through their PWWP domain (in blue) to different genomic regions (such as gene bodies or pericentric heterochromatin) [60, 68, 69], which leads to the methylation of the cytosines in the DNA of these regions (5-mC, black lollipops). On the contrary, DNA methylation valleys (DMVs) are conserved genomic regions that are normally found hypomethylated and associated with Polycomb-regulated developmental genes [64–67]. During aging, the H3K36 methylation machinery could become less efficient at maintaining the H3K36me2/3 landscape. This would lead to a relocation of de novo DNA methyltransferases from their original genomic reservoirs (which would become hypomethylated) to other non-specific regions such as DMVs (which would become hypermethylated and potentially lose their normal boundaries), with functional consequences for the tissues. This is also partially observed in patients with Sotos syndrome, where mutations in NSD1 potentially affect H3K36me2/3 patterns and accelerate the epigenetic aging clock as measured with the Horvath model [8]. Given that DNMT3B is enriched in the gene bodies of highly transcribed genes [60] and that we found these regions depleted in our differential methylation analysis, we hypothesize that the hypermethylation of DMVs could be mainly driven by DNMT3A instead. However, it is important to mention that our analysis does not discard a role of DNMT3B during epigenetic aging

categories such as “bivalent promoter” or “repressed polycomb” in the hypermethylated DMPs in Sotos and aging (Additional file 1: Figure S3A), which is also supported by higher levels of polycomb repressing complex 2 (PRC2, represented by EZH2) and H3K27me3, the mark deposited by PRC2 (Additional file 1: Figure S3B). This is also consistent with the results obtained for the epigenetic mitotic clock [39], where we observe a trend towards increased hypermethylation of Polycomb-bound regions in Sotos patients. Furthermore, it is worth mentioning that a mechanistic link between PRC2 recruitment and H3K36me3 has also been unravelled to occur via the Tudor domains of some polycomb-like proteins [70, 71].

A recent preprint has shown that loss-of-function mutations in DNMT3A, which cause Tatton-Brown-Rahman overgrowth syndrome, also lead to a higher ticking rate of the epigenetic aging clock [72]. They also report positive epigenetic age acceleration in Sotos syndrome and negative acceleration in Kabuki syndrome, consistent with our results. Furthermore, they observe a DNA methylation signature in the DNMT3A mutants characterized by widespread hypomethylation, with a modest enrichment of DMPs in the regions upstream of the transcription start site, shores, and enhancers [72], which we also detect in our “Hypo-Hypo DMPs” (those that become hypomethylated both during physiological aging and in Sotos). Therefore, the hypomethylation observed in our “Hypo-Hypo DMPs” is consistent with a reduced methylation activity of DNMT3A, which in our system could be a consequence of the decreased recruitment of DNMT3A to genomic regions that have lost H3K36 methylation (Fig. 5).

Interestingly, H3K36me3 is required for the selective binding of the de novo DNA methyltransferase DNMT3B to the bodies of highly transcribed genes [60]. Furthermore, DNMT3B loss reduces gene body methylation, which leads to intragenic spurious transcription (aka cryptic transcription) [73]. An increase in this so-called cryptic transcription seems to be a conserved feature of the aging process [56]. Therefore, the changes observed in the “Hypo-Hypo DMPs” could theoretically be a consequence of the loss of H3K36me3 and the concomitant inability of DNMT3B to be recruited to gene bodies. However, the "Hypo-Hypo DMPs" were depleted for H3K36me3, active transcription, and gene bodies when compared with the rest of the probes in the array (Fig. 3b–d), prompting us to suggest that the DNA methylation changes observed are likely mediated by DNMT3A instead (Fig. 5). Nevertheless, it is worth mentioning that the different biological replicates for the blood H3K36me3 ChIP-seq datasets were quite heterogeneous and that the absolute difference in the case of the hypomethylated Sotos DMPs, although significant due to the big sample sizes, is quite small. Thus, we cannot exclude the existence of this mechanism during human aging, and an exhaustive study on the prevalence of cryptic transcription in humans and its relation to the aging methylome should be carried out.

H3K36me3 has also been shown to guide deposition of the N6-methyladenosine mRNA modification (m^6A), an important post-transcriptional mechanism of gene regulation [74]. Interestingly, a decrease in overall m^6A during human aging has been previously reported in PBMC [75], suggesting another biological route through which an alteration of the H3K36 methylation landscape could have functional consequences for the organism.

Because of the way that the Horvath epigenetic clock was trained [8], it is likely that its constituent 353 CpG sites are a low-dimensional representation of the different genome-wide processes that are eroding the epigenome with age. Our analysis has shown that these 353 CpG sites are characterized by a higher Shannon entropy when compared with the rest of the genome, which is dramatically decreased in the case of Sotos patients. This could be related to the fact that the clock CpGs are enriched in the regions of bivalent chromatin (marked by H3K27me3 and H3K4me3), conferring a more dynamic or plastic regulatory state with levels of DNA methylation deviated from the collapsed states of 0 or 1. Interestingly, EZH2 (part of polycomb repressing complex 2, responsible for H3K27 methylation) is an interacting partner of DNMT3A and NSD1, with mutations in NSD1 affecting the genome-wide levels of H3K27me3 [76]. Furthermore, Kabuki syndrome was weakly identified in our screen as having an epigenome younger than expected, which could be related to the fact that they show postnatal dwarfism [77, 78]. Kabuki syndrome is caused by loss-of-function mutations in KMT2D [77, 78], a major mammalian H3K4 monomethyltransferase [79]. Additionally, H3K27me3 and H3K4me3 levels can affect lifespan in model organisms [3]. It will be interesting to test whether bivalent chromatin is a general feature of multi-tissue epigenetic aging clocks.

Thus, DNMT3A, NSD1, and the machinery in control of bivalent chromatin (such as EZH2 and KMT2D) contribute to an emerging picture on how the mammalian epigenome is regulated during aging, which could open new avenues for anti-aging drug development. Mutations in these proteins lead to different developmental disorders with impaired growth defects [30], with DNMT3A, NSD1, and potentially KMT2D also affecting epigenetic aging. Interestingly, EZH2 mutations (which cause Weaver syndrome, Table 1) do not seem to affect the epigenetic clock in our screen. However, this syndrome has the smallest number of samples (7), and this could limit the power to detect any changes.

Our screen has also revealed that Rett syndrome and fragile X syndrome (FXS) could potentially have an accelerated epigenetic age. It is worth noting that FXS is caused by an expansion of the CGG trinucleotide repeat located in the 5′ UTR of the *FMR1* gene [80]. Interestingly, Huntington's disease, caused by a trinucleotide repeat expansion of CAG, has also been shown to accelerate epigenetic aging of the human brain [23], pointing towards trinucleotide repeat instability as an interesting molecular mechanism to look at from an aging perspective. It is important to notice that the conclusions for Rett syndrome, FXS, and Kabuki syndrome were very dependent on the age range used in the healthy control (Additional file 1: Figure S2A), and these results must therefore be treated with caution.

Our study has several limitations that we tried to address in the best possible way. First of all, given that DNA methylation data for patients with developmental disorders is relatively rare, some of the sample sizes were quite small. It is thus possible that some of the other developmental disorders assessed are epigenetically accelerated but we lack the power to detect this. Furthermore, individuals with the disorders tend to get sampled when they are young, i.e., before reproductive age. Horvath's clock adjusts for the different rates of change in the DNA methylation levels of the clock CpGs before and after reproductive age (20 years in humans) [8], but this could still have an effect on the predictions, especially if the control is not properly age-matched. Our solution was to discard those developmental disorders with less than 5 samples, and we required them to have at least 2 samples with an age ≥ 20 years, which reduced the list of final disorders included to the ones listed in Table 1.

Future studies should increase the sample size and follow the patients during their entire lifespan in order to confirm our findings. Directly measuring the functional changes in the H3K36 methylation landscape (or its machinery) during human aging will further validate this work. Moreover, it would be interesting to identify mutations that affect, besides the mean, the variance of epigenetic age acceleration, since changes in methylation variability at single CpG sites with age have been associated with fundamental aging mechanisms [46]. Finally, testing the influence of H3K36 methylation on the epigenetic clock and lifespan in mice will provide deeper mechanistic insights.

Conclusions

The epigenetic aging clock has created a new methodological paradigm to study the aging process in humans. However, the molecular mechanisms that control its ticking rate are still mysterious. In this study, by looking at patients with developmental disorders, we have demonstrated that Sotos syndrome accelerates epigenetic aging and uncovered a potential role of the H3K36 methylation machinery as a key component of the *epigenetic maintenance system* in humans. We hope that this research will shed some light on the different processes that erode the human epigenetic landscape during aging and provide a new hypothesis about the mechanisms behind the epigenetic aging clock.

Methods

Sample collection and annotation

We collected DNA methylation data generated with the Illumina Infinium HumanMethylation450 BeadChip (450K array) from human blood. In the case of the developmental disorder samples, we combined public data with the data generated in-house for other clinical studies (Table 1, Additional file 2) [31]. We took all the data for developmental disorders that we could find in order to perform unbiased screening. The healthy samples used to build the control were mainly obtained from public sources (Additional file 3). Basic metadata (including the chronological age) was also stored. All the mutations in the developmental disorder samples were manually curated using Variant Effect Predictor [81] in the GRCh37 (hg19) human genome assembly. Those samples with a variant of unknown significance that had the characteristic DNA methylation signature of the disease were also included (they are labelled as "YES_predicted" in Additional file 2). In the case of fragile X syndrome (FXS), only male samples with full mutation (> 200 repeats) [80] were included in the final screen. As a consequence, only the samples with a clear molecular and clinical diagnosis were kept for the final screen.

Pre-processing, QC, and filtering the data for the epigenetic clock calculations

Raw DNA methylation array data (IDAT files) were processed using the *minfi* R package [82]. Raw data were background-corrected using *noob* [83] before calculating the beta values. In the case of the beta values which are input to Horvath's model, we observed that background correction did not have a major impact in the final predictions of epigenetic age acceleration in the control as long as we corrected for batch effects (Fig. 1c, Additional file 1: Figure S5A). We decided to keep the *noob* background correction step for consistency with the rest of the pipelines. Epigenetic age (*DNAmAge*) was calculated using the code from Horvath, which includes an internal normalization step against a blood gold standard [8]. The scripts are available in our GitHub repository (https://github.com/demh/epigenetic_ageing_clock) for the use of the community [84].

Quality control (QC) was performed in all samples. Following the guidelines from the *minfi* package [82], only those samples that satisfied the following criteria were kept for the analysis: the sex predicted from the DNA methylation data was the same as the reported sex in the metadata, they passed BMIQ normalization and $\frac{\text{median}(\log_2 M)+\text{median}(\log_2 U)}{2} \geq 10.5$, where M is the methylated intensity and U the unmethylated intensity for the array probes.

Correcting for batch effects

In order to correct for batch effects that could confound the conclusions from our analysis, we decided to make use of the control probes available in the 450K array. These probes capture only the technical variance in negative controls and different steps of the array protocol, such as bisulfite conversion, staining or hybridization [34, 85]. We performed PCA (with centering but not scaling using the *prcomp* function in R) on the raw intensities of the control probes (847 probes × 2 channels = 1694 intensity values) for all our controls ($N = 2218$) and cases ($N = 666$) that passed QC (Fig. 1a). Including the technical PCs as covariates in the models to calculate epigenetic age acceleration (EAA) improved the error from the predictions in the controls (Fig. 1c, Additional file 1: Figure S5A). The optimal number of PCs was found by making use of the *findElbow* function from [86].

Correcting for cell composition

The proportions of different blood cell types change with age and this can affect the methylation profiles of the samples. Therefore, when calculating the epigenetic age acceleration, it is important to compare the models with and without cell type proportions included as covariates [38]. Cell type proportions can be estimated from DNA methylation data using different deconvolution algorithms [87]. In the context of the epigenetic clock, most of the studies have used the Houseman method [88]. We have benchmarked different reference-based deconvolution strategies (combining different preprocessing steps, references, and deconvolution algorithms) against a gold standard dataset (GSE77797) [89]. Our results suggest that using the IDOL strategy [89] to build the blood reference (from the Reinius et al. dataset, GSE35069) [90], together with the Houseman algorithm [88] and some pre-processing steps (*noob* background correction, probe filtering, BMIQ normalization), leads to the best cell type proportions estimates, i.e., those that minimize the deviations between our estimates and the real cell type composition of the samples in the gold standard dataset (Additional file 1: Figure S5B, Additional file 4). We used the *epidish* function from the *EpiDISH* R package [91] for these purposes.

Calculating the epigenetic age acceleration and performing the main screen

Only those developmental disorders for which we had at least 5 samples, with 2 of them with an age ≥ 20 years, were included in the main screen ($N = 367$). Healthy samples that matched the age range of those disorders (0–55 years, $N = 1128$) were used to train the following linear models (the *control models*):

(I) Without cell composition correction (CCC):

$$\mathrm{DNAmAge} \sim \mathrm{Age} + \mathrm{Sex} + \mathrm{PC1} + \mathrm{PC2} + \ldots + \mathrm{PCN}$$

(II) With cell composition correction (CCC):

$$\begin{aligned}\mathrm{DNAmAge} \sim\ & \mathrm{Age} + \mathrm{Sex} + \mathrm{Gran} + \mathrm{CD4T} + \mathrm{CD8T} \\ & + \mathrm{B} + \mathrm{Mono} + \mathrm{NK} + \mathrm{PC1} + \mathrm{PC2} + \ldots \\ & + \mathrm{PCN}\end{aligned}$$

where *DNAmAge* is the epigenetic age calculated using Horvath's model [8], *Age* is the chronological age, *PCN* is the *N*th technical PC obtained from the control probes ($N = 17$ was the optimal, Fig. 1c) and *Gran*, *CD4T*, *CD8T*, *B*, *Mono*, and *NK* are the different proportions of the blood cell types as estimated with our deconvolution strategy. The linear models were fitted in R with the *lm* function, which uses least-squares.

The residuals from a control model represent the epigenetic age acceleration (EAA) for the different healthy samples, which should be centered around zero after batch effect correction (Additional file 1: Figure S1E, Fig. 1d). Then, the median absolute error (MAE) can be calculated as (Fig. 1c, Additional file 1: Figure S5A):

(III) $\mathrm{MAE} = \mathrm{median}(\mathit{abs}(\mathrm{EAA}i))$

where EAA_i is the epigenetic age acceleration for a healthy sample from the control.

Once the control models are established, we can calculate the EAA for the different samples with a developmental disorder (cases) by taking the difference between the epigenetic age (*DNAmAge*) for the case sample and the predicted value from the corresponding control model (with or without cell composition correction). Finally, the distributions of the EAA for the different developmental disorders were compared against the EAA distribution for the healthy controls using a two-sided Wilcoxon's test. *p* values were adjusted for multiple testing using Bonferroni correction and a significance level of $\alpha = 0.01$ was applied.

A similar approach was used in the case of the other epigenetic clocks assessed. The linear coefficients for the different probes were obtained from the original publications [9, 40, 41]. In the case of the skin-blood clock, the same age transformation employed for the Horvath's clock was applied [41]. Due to our filtering criteria, some array probes were missing, which could slightly affect the predictions of the different epigenetic clocks: Hannum's clock [9] (68/71 probes available), Lin's clock [40] (97/99 probes available), and the skin-blood clock [41] (385/391 probes available). This may be the reason behind the offset observed, particularly prominent in the predictions of Lin's clock (Additional file 1: Figure S2C-E). Nevertheless, this bias is present in both Sotos and control samples, and therefore, it is unlikely that it affects the main conclusions.

Calculating *pcgtAge* and Shannon entropy

Raw DNA methylation data (IDAT files) was background-corrected using *noob* [83]. Next, we filtered out the probes associated with SNPs, cross-reactive probes [92], and probes from the sex chromosomes, before performing BMIQ intra-array normalization to correct for the bias in probe design [93]. Then, we calculated *pcgtAge* as the average of the beta values for the probes that constitute the epigenetic mitotic clock [39]. It is worth noting that only 378 out of the 385 probes were left after our filtering criteria.

Shannon entropy was calculated as previously described [9]:

$$\text{(IV)}\ \mathrm{Entropy} = \frac{1}{N \times \log_2\left(\frac{1}{2}\right)} \times \sum_{i=1}^{N} \left[\beta_i \times \log_2(\beta_i) + (1 - \beta_i) \times \log_2(1 - \beta_i)\right]$$

where β_i represents the methylation beta value for the *i*th probe (CpG site) in the array, $N = 428{,}266$ for the genome-wide entropy, and $N = 353$ for Horvath clock sites entropy.

In order to calculate the *pcgtAge* and Shannon entropy acceleration, we followed a similar strategy to the one reported for EAA with CCC, fitting the following linear models:

(V) $\mathit{pcgtAge} \sim \mathrm{Age} + \mathrm{Sex} + \mathrm{Gran} + \mathrm{CD4T} + \mathrm{CD8T} + \mathrm{B} + \mathrm{Mono} + \mathrm{NK} + \mathrm{PC1} + \ldots + \mathrm{PC17}$

(VI) $\mathrm{Entropy} \sim \mathrm{Age} + \mathrm{Sex} + \mathrm{Gran} + \mathrm{CD4T} + \mathrm{CD8T} + \mathrm{B} + \mathrm{Mono} + \mathrm{NK} + \mathrm{PC1} + \ldots + \mathrm{PC17}$

It is worth mentioning that we observed a remarkable effect of the batch on the Shannon entropy calculations, which generated high entropy variability for a given age (Additional file 1: Figure S4C,D). Thus, accounting for technical variation becomes crucial when assessing this type of data, even after background correction, probe filtering, and BMIQ normalization.

Identifying differentially methylated positions

DMPs were identified using a modified version of the *dmpFinder* function in the *minfi* R package [82], where we accounted for other covariates. The aging DMPs (aDMPs) were calculated using the control samples that were included in the screen (age range 0–55 years, $N = 1128$) and the following linear model (p values and regression coefficients were extracted for the *Age* covariate):

$$\text{(VII)}\ \beta_i \sim \text{Age} + \text{Sex} + \text{Gran} + \text{CD4T} + \text{CD8T} + \text{B} + \text{Mono} + \text{NK} + \text{PC1} + \ldots + \text{PC17}$$

where β_i represents the methylation beta value for the ith probe (CpG site) in the array.

The Sotos DMPs were calculated by comparing the Sotos samples ($N = 20$) against the control samples ($N = 51$) from the same dataset (GSE74432) [44] using the following linear model (p values and regression coefficients were extracted for the *Disease_status* covariate):

$$\text{(VIII)}\ \beta_i \sim \text{Disease\_status} + \text{Age} + \text{Sex} + \text{Gran} + \text{CD4T} + \text{CD8T} + \text{B} + \text{Mono} + \text{NK} + \text{PC1} + \ldots + \text{PC17}$$

We selected as our final DMPs those CpG probes that survived our analysis after Bonferroni multiple testing correction with a significance level of $\alpha = 0.01$.

(Epi) genomic annotation of the CpG sites

Different (epi) genomic features were extracted for the CpG sites of interest. All the data were mapped to the *hg19* assembly of the human genome.

The continuous features were calculated by extracting the mean value in a window of ± 200 bp from the CpG site coordinate using the *pyBigWig* package [94]. We chose this window value based on the methylation correlation observed between neighboring CpG sites in previous studies [95]. The continuous features included (Additional file 5) the following:

- ChIP-seq data from ENCODE (histone modifications from peripheral blood mononuclear cells or PBMC; EZH2, as a marker of polycomb repressing complex 2 binding, from B cells; RNF2, as a marker of polycomb repressing complex 1 binding, from the K562 cell line). We obtained Z-scores (using the *scale* function in R) for the values of "fold change over control" as calculated in ENCODE [96]. When needed, biological replicates of the same feature were aggregated by taking the mean of the Z-scores in order to obtain the "normalised fold change" (NFC).
- ChIP-seq data for LaminB1 (GSM1289416, quantified as "normalised read counts" or NRC) and Repli-seq data for replication timing (GSM923447, quantified as "wavelet-transformed signals" or WTS). We used the same data from the IMR90 cell line as in [97].
- Total RNA-seq data (rRNA depleted, from PBMC) from ENCODE. We calculated Z-scores after aggregating the "signal of unique reads" (*sur*) for both strands (+ and −) in the following manner:

$$\text{(IX)}\ \text{RNA}_i = \log_2(1 + \text{sur}_{i+} + \text{sur}_{i-})$$

where RNA_i represents the RNA signal (that then needs to be scaled to obtain the "normalised RNA expression" or NRE) for the ith CpG site.

The categorical features were obtained by looking at the overlap (using the *pybedtools* package) [98] of the CpG sites with the following:

- Gene bodies, from protein-coding genes as defined in the basic gene annotation of GENCODE release 29 [99].
- CpG islands (CGIs) were obtained from the UCSC Genome Browser [100]. Shores were defined as regions 0 to 2 kb away from CGIs in both directions and shelves as regions 2 to 4 kb away from CGIs in both directions as previously described [95, 101].
- Chromatin states were obtained from the K562 cell line in the Roadmap Epigenomics Project (based on imputed data, 25 states, 12 marks) [102]. A visualization for the association between chromatin marks and chromatin states can be found in [103]. When needed for visualization purposes, the 25 states were manually collapsed to a lower number of them.

We compared the different genomic features for each one of our subsets of CpG sites (hypomethylated aDMPs, hypomethylated Sotos DMPs) against a control set. This control set was composed of all the probes from the background set from which we removed the subset that we were testing. In the case of the comparisons against the 353 Horvath clock CpG sites, a background set of the 21,368 (21K) CpG probes used to train the original Horvath model [8] was used. In the case of the genome-wide comparisons for aging and Sotos syndrome, a background set containing all 428,266 probes that passed our pre-processing pipeline (450K) was used.

The distributions of the scores from the continuous features were compared using a two-sided Wilcoxon's test. In the case of the categorical features, we tested for enrichment using Fisher's exact test.

Differences in the clock CpGs beta values for Sotos syndrome

To compare the beta values of the Horvath clock CpG sites between our healthy samples and Sotos samples, we fitted the following linear models in the healthy samples (*control CpG models*, Additional file 1: Figure S3C, Additional file 6):

$$(X)\ \beta_i \sim \mathrm{Age} + \mathrm{Age}^2 + \mathrm{Sex} + \mathrm{Gran} + \mathrm{CD4T} + \mathrm{CD8T} + \mathrm{B} + \mathrm{Mono} + \mathrm{NK} + \mathrm{PC1} + \ldots + \mathrm{PC17}$$

where β_i represents the methylation beta values for the ith probe (CpG site) in the 353 CpG clock sites. The Age^2 term allows accounting for non-linear relationships between chronological age and the beta values.

Finally, we calculated the difference between the beta values in Sotos samples and the predictions from the *control CpG models* and displayed these differences in an annotated heatmap (Additional file 1: Figure S3D).

Code availability

All the code used to perform the analyses here presented can be found in our GitHub repository (https://github.com/demh/epigenetic_ageing_clock) under GNU General Public License v3.0 [84].

Additional files

Additional file 1: Supplementary figures that complement the main manuscript. (PDF 2877 kb)

Additional file 2: Information for the samples with developmental disorders (cases) that were included in the main screen ($N = 367$). (TSV 216 kb)

Additional file 3: Information for the healthy control samples that were included in the main screen ($N = 1128$). (TSV 633 kb)

Additional file 4: Information about the different blood cell type deconvolution strategies that were benchmarked against the gold standard dataset (GSE77797). (XLSX 13 kb)

Additional file 5: Information (including the source) about the continuous (epi) genomic features (ChIP-seq and RNA-seq data) that were included in our analysis to annotate the different sets of CpG sites. (CSV 1 kb)

Additional file 6: DNA methylation (beta value) profiles for the 353 Horvath's epigenetic clock CpG sites during aging for healthy individuals (gray) and Sotos patients (orange). A linear model (displayed in dark gray) can be fixed to each CpG site to model the changes in beta value with chronological age in the controls (gray). Information about whether the site is a differentially methylated position during aging (aDMP) or in Sotos patients (Sotos DMP) is also provided. Hyper, hypermethylated; Hypo, hypomethylated; No, not statistically significant after Bonferroni correction. (PDF 2811 kb)

Additional file 7: Review history. (DOCX 42 kb)

Abbreviations

450K array: Illumina Infinium HumanMethylation450 BeadChip; aDMPs: Differentially methylated positions during aging; ASD: Autism spectrum disorder; ATR-X: Alpha thalassemia/mental retardation X-linked syndrome; CCC: Cell composition correction; DMPs: Differentially methylated positions; EAA: Epigenetic age acceleration; FXS: Fragile X syndrome; IEAA: Intrinsic epigenetic age acceleration; iPSCs: Induced pluripotent stem cells; MAE: Median absolute error; PBMC: Peripheral blood mononuclear cells; PCA: Principal component analysis; PCs: Principal components; rRNA: Ribosomal RNA; UTR: Untranslated region

Acknowledgements

We thank all the members of the Reik and Thornton Laboratories for the helpful discussions. Specifically, we would like to acknowledge Melike Dönertas, Dr. Gos Micklem, and Dr. Judith Zaugg for their advice during the last years. Furthermore, we thank all the authors and groups that answered our emails, allowed us to use the raw DNA methylation data that they generated, and provided missing metadata, including Dr. Kimberly Aldinger, Dr. Sebastian Morán, Dr. Manuel Castro, Dr. Juan Sandoval, Dr. Pankaj Chopra, Dr. Akdes Serin-Harmanci, Dr. Weng Khong Lim, Dr. Reiner Schulz, Tina Wang, Dr. Barry Demchack, and Professor Trey Ideker. Finally, we would like to thank Parvathi "Ale" Subbiah for designing the figure displaying our current working model.

Review history

The review history is available at Additional file 7.

Authors' contributions

DEMH, TMS, WR, and JMT designed the study. DEMH, EAE, and MJB conducted the data analysis. EAE, SC, RW, and BS generated part of the DNA methylation dataset. MJB and OS provided crucial statistical input. DEMH, WR, and JMT interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Funding

EMBL predoctoral fellowship (to DEMH); fellowship from the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant number 664726; to MJB). WR acknowledges funding from BBSRC. JMT acknowledges funding from EMBL.

Availability of data and materials

Part of the DNA methylation data and metadata was obtained from the GEO public repository and are available under the following accession numbers: GSE104812 [104], GSE111629 [105], GSE116300 [106], GSE35069 (to build the reference for cell composition estimation) [107], GSE40279 [108], GSE41273 [109], GSE42861 [110], GSE51032 [111], GSE55491 [112], GSE59065 [113], GSE61496 [114], GSE74432 [115], GSE77797 (gold-standard for cell composition estimation) [116], GSE81961 [117], and GSE97362 [118]. The rest of the raw DNA methylation data (Europe, Feb_2016, Jun_2015, Mar_2014, May_2015, May_2016, Nov_2015, Oct_2014) are not publicly available at the time of the study as part of the conditions of the research ethical approval of the study. All the code used to perform the analyses here presented can be found in the following GitHub repository (https://github.com/demh/epigenetic_ageing_clock) under the GNU General Public License v3.0 [84].

Ethics approval and consent to participate

The study protocol has been approved by the Western University Research Ethics Board (REB ID 106302) and McMaster University and the Hamilton Integrated Research Ethics Boards (REB ID 13-653-T). All of the participants provided informed consent prior to sample collection. All of the samples and records were de-identified before any experimental or analytical procedures. The research was conducted in accordance with all relevant ethical regulations. All experimental methods comply with the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

DEMH and TMS are founders and shareholders of Chronomics Limited, a UK-based company that provides epigenetic testing. WR is a consultant and shareholder of Cambridge Epigenetix. All other authors declare that they have no competing interests.

Author details

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [2]Chronomics Ltd., Cambridge, UK. [3]Department of Pathology and Laboratory Medicine, Western University, London, Canada. [4]Molecular Genetics Laboratory, Molecular Diagnostics Division, London Health Sciences Centre, London, Canada. [5]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [6]Genetics and Genome Biology Program, Research Institute, The Hospital for Sick Children, Toronto, Canada. [7]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [8]Epigenetics Programme, The Babraham Institute, Cambridge, UK. [9]Centre for Trophoblast Research, University of Cambridge, Cambridge, UK. [10]Wellcome Sanger Institute, Hinxton, Cambridge, UK.

Received: 8 February 2019 Accepted: 3 July 2019
Published online: 14 August 2019

References

1. Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153:1194–217.
2. Benayoun BA, Pollina EA, Brunet A. Epigenetic regulation of ageing: linking environmental inputs to genomic stability. Nat Rev Mol Cell Biol. 2015;16: 593–610.
3. Sen P, Shah PP, Nativio R, Berger SL. Epigenetic mechanisms of longevity and aging. Cell. 2016;166:822–39.
4. Pal S, Tyler JK. Epigenetics and aging. Sci Adv. 2016;2:e1600584.
5. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19:371–84.
6. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA methylation clocks in aging: categories, causes, and consequences. Mol Cell. 2018;71:882–95.
7. Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. Aging (Albany NY). 2011;3:1018–27.
8. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:3156.
9. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.
10. Stubbs TM, Bonder MJ, Stark A-K, Krueger F, von Meyenn F, Stegle O, et al. Multi-tissue DNA methylation age predictor in mouse. Genome Biol. 2017;18:68.
11. Petkovich DA, Podolskiy DI, Lobanov AV, Lee S-G, Miller RA, Gladyshev VN. Using DNA methylation profiling to evaluate biological age and longevity interventions. Cell Metab. 2017;25:954–960.e6.
12. Wang T, Tsui B, Kreisberg JF, Robertson NA, Gross AM, Yu MK, et al. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. Genome Biol. 2017;18:57.
13. Thompson MJ, Chwiałkowska K, Rubbi L, Lusis AJ, Davis RC, Srivastava A, et al. A multi-tissue full lifespan epigenetic clock for mice. Aging (Albany NY). 2018;10:2832–54.
14. Meer MV, Podolskiy DI, Tyshkovskiy A, Gladyshev VN. A whole lifespan mouse multi-tissue DNA methylation clock. Elife. 2018;7:e40675.
15. Thompson MJ, von Holdt B, Horvath S, Pellegrini M. An epigenetic aging clock for dogs and wolves. Aging (Albany NY). 2017;9:1055–68.
16. Polanowski AM, Robbins J, Chandler D, Jarman SN. Epigenetic estimation of age in humpback whales. Mol Ecol Resour. 2014;14:976–87.
17. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16:25.
18. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016;8:1844–65.
19. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis. 2015;212:1563–73.
20. Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, et al. Accelerated epigenetic aging in Down syndrome. Aging Cell. 2015;14:491–5.
21. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schönfels W, Ahrens M, Heits N, Bell JT, Tsai PC, Spector TD, Deloukas P, Siebert R, Sipos B, Becker T, Röcken C, Schafmayer C, Hampe J. Obesity accelerates epigenetic aging. Proc Natl Acad Sci. 2014;111(43):15538–15543. https://doi.org/10.1073/pnas.1412759111.
22. Maierhofer A, Flunkert J, Oshima J, Martin GM, Haaf T, Horvath S. Accelerated epigenetic aging in Werner syndrome. Aging (Albany NY). 2017; 9:1143–52.
23. Horvath S, Langfelder P, Kwak S, Aaronson J, Rosinski J, Vogt TF, et al. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. Aging (Albany NY). 2016;8:1485–512.
24. Walker RF, Liu JS, Peters BA, Ritz BR, Wu T, Ophoff RA, et al. Epigenetic age analysis of children who seem to evade aging. Aging (Albany NY). 2015;7:334–9.
25. Cole JJ, Robertson NA, Rather MI, Thomson JP, McBryan T, Sproul D, et al. Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. Genome Biol. 2017;18:58.
26. Rando TA, Chang HY. Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock. Cell. 2012;148:46–57.
27. Olova N, Simpson DJ, Marioni RE, Chandra T. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. Aging Cell. 2019;18:e12877.
28. Lu AT, Xue L, Salfati EL, Chen BH, Ferrucci L, Levy D, et al. GWAS of epigenetic aging rates in blood reveals a critical role for TERT. Nat Commun. 2018;9:387.
29. Lu AT, Hannon E, Levine ME, Hao K, Crimmins EM, Lunnon K, et al. Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum. Nat Commun. 2016;7:10561.
30. Bjornsson HT. The Mendelian disorders of the epigenetic machinery. Genome Res. 2015;25:1473–81.
31. Aref-Eshghi E, Rodenhiser DI, Schenkel LC, Lin H, Skinner C, Ainsworth P, et al. Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes. Am J Hum Genet. 2018;102:156–74.
32. Hoshino A, Horvath S, Sridhar A, Chitsazan A, Reh TA. Synchrony and asynchrony between an epigenetic clock and developmental timing. Sci Rep. 2019;9:3770.
33. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13:539–52.
34. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15:503.
35. Maksimovic J, Oshlack A, Gagnon-Bartsch JA, Speed TP. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. Nucleic Acids Res. 2015;43:e106.
36. El Khoury LY, Gorrie-Stone T, Smart M, Hughes A, Bao Y, Andrayas A, et al. Properties of the epigenetic clock and age acceleration. bioRxiv. 2018:363143.
37. Marioni RE, Deary IJ, Relton CL, Suderman M, Ferrucci L, Chen BH, et al. Tracking the epigenetic clock across the human life course: a meta-analysis of longitudinal cohort data. J Gerontol Ser A. 2018;74:57–61.
38. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17:171.
39. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, et al. Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. 2016;17:205.
40. Lin Q, Weidner CI, Costa IG, Marioni RE, Ferreira MRP, Deary IJ, et al. DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. Aging (Albany NY). 2016;8:394–401.
41. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies. Aging (Albany NY). 2018;10:1758–75.
42. Leventopoulos G, Kitsiou-Tzeli S, Kritikos K, Psoni S, Mavrou A, Kanavakis E, et al. A clinical study of Sotos syndrome patients with review of the literature. Pediatr Neurol. 2009;40:357–64.
43. Kurotaki N, Imaizumi K, Harada N, Masuno M, Kondoh T, Nagai T, et al. Haploinsufficiency of NSD1 causes Sotos syndrome. Nat Genet. 2002;30:365–6.
44. Choufani S, Cytrynbaum C, Chung BHY, Turinsky AL, Grafodatskaya D, Chen YA, et al. NSD1 mutations generate a genome-wide DNA methylation signature. Nat Commun. 2015;6:10207.
45. Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. Aging (Albany NY). 2018;10:3541–57.
46. Slieker RC, van Iterson M, Luijk R, Beekman M, Zhernakova DV, Moed MH, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol. 2016;17:191.
47. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. Epigenetics Chromatin. 2018;11:25.

48. Rinaldi L, Datta D, Serrat J, Morey L, Solanas G, Avgustinova A, et al. Dnmt3a and Dnmt3b associate with enhancers to regulate human epidermal stem cell homeostasis. Cell Stem Cell. 2016;19:491–501.
49. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125:315–26.
50. Bernhart SH, Kretzmer H, Holdt LM, Jühling F, Ammerpohl O, Bergmann AK, et al. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. Sci Rep. 2016;6:37393.
51. Horvath S. DNAmAge online calculator: https://dnamage.genetics.ucla.edu/home. 2013. https://dnamage.genetics.ucla.edu/home.
52. Wagner EJ, Carpenter PB. Understanding the language of Lys36 methylation at histone H3. Nat Rev Mol Cell Biol. 2012;13:115–26.
53. Luscan A, Laurendeau I, Malan V, Francannet C, Odent S, Giuliano F, et al. Mutations in SETD2 cause a novel overgrowth condition. J Med Genet. 2014;51:512–7.
54. McDaniel SL, Hepperla AJ, Huang J, Dronamraju R, Adams AT, Kulkarni VG, et al. H3K36 methylation regulates nutrient stress response in Saccharomyces cerevisiae by enforcing transcriptional fidelity. Cell Rep. 2017;19:2371–82.
55. Ni Z, Ebata A, Alipanahiramandi E, Lee SS. Two SET domain containing genes link epigenetic changes and aging in Caenorhabditis elegans. Aging Cell. 2012;11:315–25.
56. Sen P, Dang W, Donahue G, Dai J, Dorsey J, Cao X, et al. H3K36 methylation promotes longevity by enhancing transcriptional fidelity. Genes Dev. 2015; 29:1362–76.
57. Pu M, Ni Z, Wang M, Wang X, Wood JG, Helfand SL, et al. Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes Dev. 2015;29:718–31.
58. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517:321–6.
59. Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, et al. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J Biol Chem. 2010;285:26114–20.
60. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015;520:243–7.
61. Heyn P, Logan CV, Fluteau A, Challis RC, Auchynnikava T, Martin C-A, et al. Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. Nat Genet. 2019;51:96–105.
62. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20:434–9.
63. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20:440–6.
64. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013;153:1134–48.
65. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Elife. 2013;2:e00348.
66. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet. 2013;46:17–23.
67. Li Y, Zheng H, Wang Q, Zhou C, Wei L, Liu X, et al. Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. Genome Biol. 2018;19:18.
68. Chantalat S, Depaux A, Héry P, Barral S, Thuret JY, Dimitrov S, et al. Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. Genome Res. 2011;21:1426–37.
69. Chen T, Tsujimoto N, Li E. The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. Mol Cell Biol. 2004;24:9048–58.
70. Cai L, Rothbart SB, Lu R, Xu B, Chen W-Y, Tripathy A, et al. An H3K36 methylation-engaging Tudor motif of Polycomb-like proteins mediates PRC2 complex targeting. Mol Cell. 2013;49:571–82.
71. Li H, Liefke R, Jiang J, Kurland JV, Tian W, Deng P, et al. Polycomb-like proteins link the PRC2 complex to CpG islands. Nature. 2017;549:287–91.
72. Jeffries AR, Maroofian R, Salter CG, Chioza BA, Cross HE, Patton MA, et al. Growth disrupting mutations in epigenetic regulatory molecules are associated with abnormalities of epigenetic aging. bioRxiv. 2018:477356.
73. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017;543:72–7.
74. Huang H, Weng H, Zhou K, Wu T, Zhao BS, Sun M, et al. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature. 2019;567:414–9.
75. Min K-W, Zealy RW, Davila S, Fomin M, Cummings JC, Makowsky D, et al. Profiling of m6A RNA modifications identified an age-associated regulation of AGO2 mRNA stability. Aging Cell. 2018;17:e12753.
76. Streubel G, Watson A, Jammula SG, Scelfo A, Fitzpatrick DJ, Oliviero G, et al. The H3K36me2 methyltransferase Nsd1 demarcates PRC2-mediated H3K27me2 and H3K27me3 domains in embryonic stem cells. Mol Cell. 2018; 70:371–379.e5.
77. Butcher DT, Cytrynbaum C, Turinsky AL, Siu MT, Inbar-Feigenberg M, Mendoza-Londono R, et al. CHARGE and Kabuki syndromes: gene-specific DNA methylation signatures identify epigenetic mechanisms linking these clinically overlapping conditions. Am J Hum Genet. 2017;100:773–88.
78. Aref-Eshghi E, Schenkel LC, Lin H, Skinner C, Ainsworth P, Paré G, et al. The defining DNA methylation signature of Kabuki syndrome enables functional assessment of genetic variants of unknown clinical significance. Epigenetics. 2017;12:923–33.
79. Froimchuk E, Jang Y, Ge K. Histone H3 lysine 4 methyltransferase KMT2D. Gene. 2017;627:337–42.
80. Schenkel LC, Schwartz C, Skinner C, Rodenhiser DI, Ainsworth PJ, Pare G, et al. Clinical validation of fragile X syndrome screening by DNA methylation array. J Mol Diagnostics. 2016;18:834–41.
81. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
82. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30: 1363–9.
83. Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013;41:e90.
84. Martin-Herranz DE. demh/epigenetic_ageing_clock: Epigenetic ageing clock v1.1.0. GitHub repository: https://github.com/demh/epigenetic_ageing_clock/. 2019. doi:https://doi.org/10.5281/zenodo.3263907.
85. Illumina. GenomeStudio® methylation module v1.8 User Guide. 2010.
86. Akalin A. AmpliconBiSeq GitHub repository: findElbow function; 2014.
87. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9:757–68.
88. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
89. Koestler DC, Jones MJ, Usset J, Christensen BC, Butler RA, Kobor MS, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics. 2016;17:120.
90. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7:e41361.
91. Teschendorff AE, Zheng SC. EpiDISH bioconductor package. 2017. https://bioconductor.org/packages/release/bioc/html/EpiDISH.html.
92. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.
93. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics. 2012;29:189–96.
94. Richter AS, Ryan DP, Kilpert F, Ramírez F, Heyne S, Manke T. pyBigWig GitHub Repository. https://github.com/deeptools/pyBigWig.
95. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16:14.
96. Consortium TEP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
97. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018;50:591–602.

98. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27:3423–4.
99. Frankish A, Bignell A, Berry A, Yates A, Parker A, Schmitt BM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2018;47:D766–73.
100. Bock C, Walter J, Paulsen M, Lengauer T. CpG island mapping by epigenome prediction. PLoS Comput Biol. 2007;3:e110.
101. Martin-Herranz DE, Ribeiro AJM, Krueger F, Thornton JM, Reik W, Stubbs TM. cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches. Nucleic Acids Res. 2017;45:11559–69.
102. Consortium NREM. Roadmap epigenomics chromatin state model: raw data. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/catMat/hg19_chromHMM_imputed25.gz. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/catMat/hg19_chromHMM_imputed25.gz.
103. Consortium NREM. Roadmap epigenomics chromatin state model: emission parameters. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/emissions_25_imputed12marks.png. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/emissions_25_imputed12marks.png.
104. Wang Z, Shi L. Epigenome analysis of whole blood samples in Chinese children. GSE104812. Gene Expression Omnibus. 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104812.
105. Ritz B, Horvath S. Genome wide DNA methylation study of Parkinson's disease in whole blood samples. GSE111629. Gene Expression Omnibus. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111629.
106. Brucato M, Sobreira N, Zhang L, Ladd-Acosta C, Ongaco C, Romm J, et al. Patients with a Kabuki syndrome phenotype demonstrate DNA methylation abnormalities. GSE116300. Gene Expression Omnibus. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116300.
107. Reinius L, Acevedo N, Joerink M, Pershagen G, Dahlén S, Greco D, et al. Differential DNA methylation in purified human blood cells. GSE35069. Gene Expression Omnibus. 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35069.
108. Zhang K, Ideker T. Genome-wide methylation profiles reveal quantitative views of human aging rates. GSE40279. Gene Expression Omnibus. 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279.
109. Warren S, Chopra P. Genome-wide analysis identifies aberrant methylation in Fragile X syndrome is specific to the FMR1 locus. GSE41273. Gene Expression Omnibus. 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41273.
110. Liu Y, Feinberg A. Differential DNA methylation in rheumatoid arthritis. GSE42861. Gene Expression Omnibus. 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861.
111. Polidoro S, Campanella G, Krogh V, Palli D, Panico S, Tumino R, et al. EPIC-Italy at HuGeF. GSE51032. Gene Expression Omnibus. 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51032.
112. Prickett A, Ishida M, Böhm S, Frost J, Puszyk W, Abu-Amero S, et al. Genomewide methylation analysis in Silver Russell syndrome patients. GSE55491. Gene Expression Omnibus. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55491.
113. Milani L, Peterson P. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. GSE59065. Gene Expression Omnibus. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59065.
114. Tan Q, Christiansen L, Frost M. Epigenetic signature of birth-weight discordance in Danish twins. GSE61496. Gene Expression Omnibus. 2014. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61496.
115. Choufani S, Turinsky A, Weksberg R. NSD1 mutations generate a genome-wide DNA methylation signature. GSE74432. Gene Expression Omnibus. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74432.
116. Koestler D, Christensen B, Wiencke J, Kelsey K. DNA methylation profiling of whole blood and reconstructed mixtures of purified leukocytes isolated from human adult blood. GSE77797. Gene Expression Omnibus. 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77797.
117. Li Yim A, Duijvis N, Zhao J, de Jonge W, D'Haens G, Mannens M, et al. Peripheral blood methylation profiling of female Crohn's disease patients. GSE81961. Gene Expression Omnibus. 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81961.
118. Butcher D, Choufani S, Turinsky A, Weksberg R. CHARGE and Kabuki syndromes: gene-specific DNA methylation signatures. GSE97362. Gene Expression Omnibus. 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97362.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.