Epigenetic fidelity in complex biological systems and implications for ageing

Thomas Duffield[1], Laura Csuka[2], Arda Akalan[1], Gustavo Vega Magdaleno[1], Daniel Palmer[3], and João Pedro de Magalhães[4]

[1] *Department of Musculoskeletal and Ageing Science, University of Liverpool*
[2] *Department of Computer Science, University of Oxford*
[3] *Institute for Biostatistics and Computer Science in Medicine and Ageing Research, University of Rostock*
[4] *Genomics of Ageing and Rejuvenation Lab, Institute of Inflammation and Ageing, University of Birmingham*

Abstract

The study of age is plagued by a lack of delineation between the causes and effects within the ageing phenotype. This has made it difficult to fully explain the biological ageing process from first principles with a single definition. Lacking a clear description of the underlying root cause of biological age confounds clarity in this critical field. In this paper, we demonstrate that the epigenetic system has a built-in, unavoidable fidelity limitation and consequently demonstrate that there is a distinct class of DNA methylation loci that increases in variance in a manner tightly correlated with chronological age. We demonstrate the existence of epigenetic 'activation functions' and that topological features beyond these activation functions represent deregulation. We show that the measurement of epigenetic fidelity is an accurate predictor of cross-species age and present a deep-learning model that predicts exclusively from knowledge of variance. We find that the classes of epigenetic loci in which variation correlates with chronological age control genes that regulate transcription and suggest that the inevitable consequence of this is a feedback cycle of system-wide deregulation causing a progressive collapse into the phenotype of age. This paper represents a novel theory of biological systemic ageing with arguments as to why, how and when epigenetic ageing is inevitable.

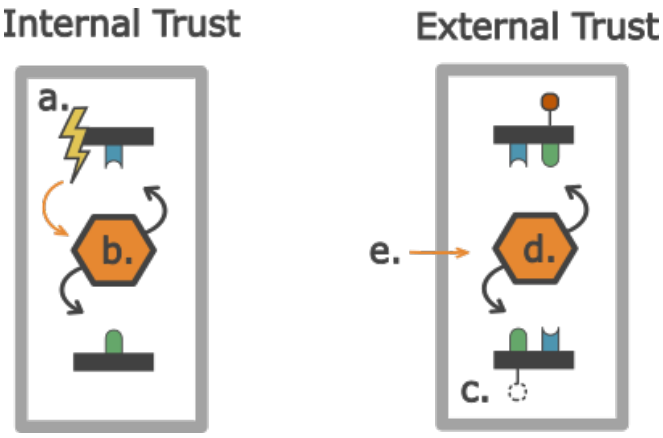
Introduction

Despite increased research and the undeniable importance and impact of ageing in medicine and society (1), the exact nature of human ageing and its causative mechanisms remain largely controversial. Many theories have been put forward attempting to explain the ageing process (2), yet the underlying molecular drivers of the human ageing process continue to be a subject of great interest and intense debate (3).

Recent studies have put the limelight on the potential role of epigenetic modifications in ageing (4; 5; 6). These include the discovery of epigenetic clocks, highly accurate predictors of chronological age, based on a relatively small number of methylation sites (7; 8). Epigenetic clocks are associated with mortality, they can predict chronological age from various tissues, across the lifespan and in multiple species, although their mechanistic basis remains the subject of debate (5; 9). In addition, multiple changes in methylation and other epigenetic modifications have been reported with age, both in human and animal models (4; 6; 10).

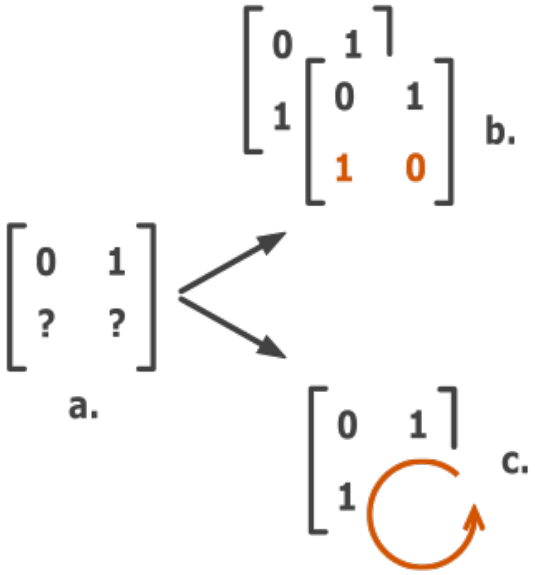
It has been proposed that epigenetic changes are causative in ageing (5), and a recent study has suggested that DNA damage response-induced loss of epigenetic information drives ageing (11). More broadly, the information theory of ageing has suggested that loss of epigenetic information with age is a major driver of the ageing process (12; 11). It has also been suggested that pre-programmed shifts in epigenetic information states with age are a major determinant of ageing phenotypes (13). As such, understanding the basis of epigenetic clocks, and how epigenetic changes could impact ageing is a major and important open question. Moreover, despite efforts to understand the informatic character of ageing, there has been comparatively little research on what makes mammalian ageing inevitable.

In this work, we develop a conceptual model to explain the ageing process based on first principles. We demonstrate that the epigenetic system has unique inherent informatic properties that progressively acquire informatic corruption, meaning that with age epigenetic information fidelity cannot be maintained. Our model is further supported by empirical data from humans and other species, and we derive a predictor of age based solely on measures of epigenetic variation.

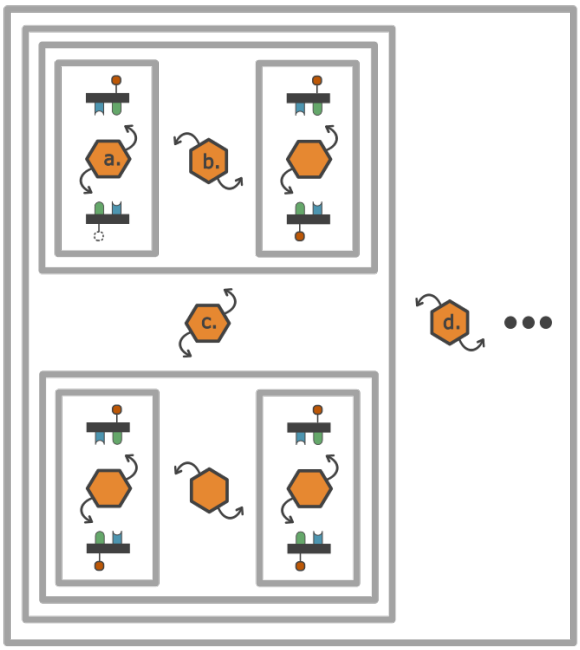


(i) *Internal and external trust information*

(ii) *Biological damage is information erasure*



(iii) *Reconstitution of information*



(iv) *Mirror backups have an infinitely regressive outer shell*

Figure 1: *i. Natural selection infers that in single-strand breaks the side with the broken backbone (a.) is most likely to contain the incorrect base. A system for deducing which strand to use as a mirror backup (b.) will have access to this information. Methylation damage (c.) cannot provide this information, so (d.) would require information from an external scope (e.) to make the same comparison. ii. Incorrect modification of CpG methylation can be thought of as information erasure, removing part of the state information that allows for the recovery of the original state. iii. Epigenetic damage is information loss (a.) that requires repair either with a mirror backup from which to duplicate information (b.) or an algorithm with which to define it according to original principles(c.). iv. Epigenetic mirrored backups represent a vicious infinite regress of endlessly nesting scopes. Comparing two strands requires a system of trust recognition (a.) that, if subject to noise, would itself require a mirror backup. These two systems would themselves require an external system of trust (b.), which can itself make errors, requiring a backup and another system of trust (c.) and so on (d.)*

1 The information fidelity theory of age

Repairing Damaged Information

Any state, including that of DNA methylation, can be thought of as a state of information(Fig 1(ii)), and therefore epigenetic damage (which we define as any epigenetic change that reduces the organism's overall chance of survival, and thus is selected towards system ontology) represents information loss. When information is destroyed, there are only two possible mechanisms by which it can be recovered(Fig 1(iii)). Information in the original state can be recovered from an identical backup, via a system encoded to know which is which. Alternatively, information can be reconstituted through an algorithm: a series of rules that defined the original information state. All of these systems of data recovery must be applied consequently to damage through either observation or prediction of data loss.

Mirror backup is impossible

To create a mirror backup for DNA methylation, an object with identical informatic properties would have to exist from which to duplicate the information, which would also have to behave in an identical manner to the first in response to noise and errors. Were this not the case, there would no longer be one-to-one parity between original and backup, and the process of comparison would itself become noisy. A system would then be necessary to duplicate changes between original and backup, maintaining parity, encoding trust, and indicating which of the two DNA methylation signals should be treated as a backup in the event of damage. In a biological context, such a system would itself inevitably be subject to error, allowing noise to enter the decision-making governing trust and therefore requiring another mechanism for mitigation and correction. In essence, just as DNA methylation is the single outer layer of control for DNA, DNA methylation itself would require the same system, which would be subject to the exact problems it was intended to avoid (Fig 1(iv)). This "nesting doll problem" is infinitely recursive: it is logically impossible in a noise-filled environment to design a signal without a component in which all damage has a mirror backup. We suggest a flawless epigenetic mirror is impossible as an example of vicious infinite regress (14), extremely similar to Bradley's regress (**?** 15). Although described here in terms of individual methylation loci, this process holds true for regions of methylation or even systems of comparison between chromosomes. Any such comparison requires a 'comparer', which becomes the point of entry for signal corruption, unless it itself has a backup and so on.

Algorithmic fidelity is restricted

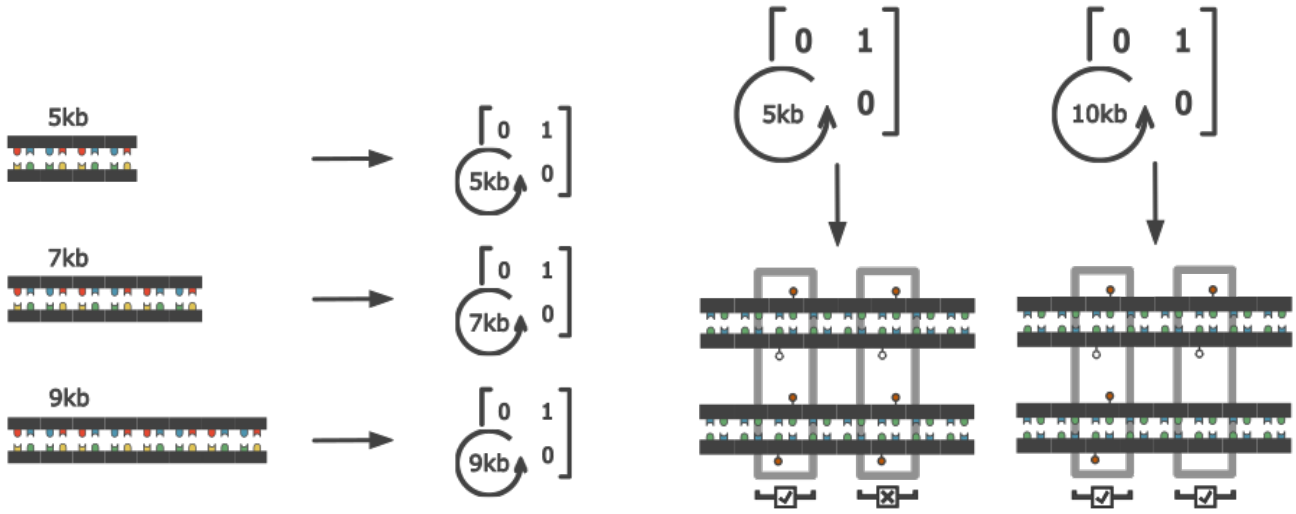
Lacking a mirror backup, any information lost in epigenetic damage must be reconstituted using some form of algorithm. Any algorithm that reconstitutes information must itself be encoded which, in the context of the cell, means genetically encoded in DNA. This has a consequent cost to the cell (for example, the more DNA used, the greater the chance of mutation), meaning that any increase in survival cost must be offset with additional functionality. Minimum algorithm size increases with the complexity of information it is to define: an increase in the latter must result in an increase in the former(Fig 2(i)). This means algorithm size is also related to the fidelity by which it reconstitutes lost information because low fidelity reconstitution represents a reduction in information from the original (Fig 2(ii)), essentially performing lossy compression (16). A perfect reconstitution requires the exclusive use of lossless compression and has consequently higher storage requirements. Natural selection will not select for lossless compression if the cost of the additional information outweighs the benefit to survival, meaning in all cases one should expect DNA compression to be lossy (except in the case of individual errors with infinite cost to survival, e.g. errors leading to cancer). With DNA methylation containing two legal character states, defining it with perfect fidelity would be equivalent to binary key definition in cryptology, becoming exponentially large as regions contain more CpG loci. With approximately 20 million CpG in the human genome, perfect fidelity is therefore impossible. Even working under the assumption that epigenetic regions represent the states to define, there are over 20000 CpG islands in the human genome and an uncountable number of cellular identities to define.

This is not to say that reconstruction is generally impossible, but that high-fidelity reconstruction is extremely informatically expensive and impossible to perform over a large number of cellular states.

As a result, an epigenetic algorithm reconstituting lost information would be forced to work within a spectrum between total lack of fidelity (randomly recreating data) and flawless fidelity, with the massive amount of information required for high fidelity restricting the majority of systems to error-prone reconstitution of damage.

Legal Characters and Trust

When damage occurs in DNA it almost always produces a dictionary illegal character on one of the strands. In such cases (e.g., bulky adducts) DNA repair mechanisms can cheaply and effectively recognise that these new 'characters' in the DNA signal fall outside of the pre-defined set of legal dictionary characters: A, T, C, G. Both in this situation and when a dictionary legal character is created, DNA repair mechanisms must look for more information to determine which of the two strands to treat



(i) *Dna cost impacts repair fidelity* **(ii)** *Repair fidelity impacts error retention*

Figure 2: *i. Any algorithm that corrects for epigenetic error must be comprised of proteins or RNA, which requires a region of DNA to encode. The size of the genes encoding the system will be related to the sophistication of the algorithm, but this will come at a survival cost commensurate to size. ii. The more sophisticated the algorithm for correcting error, the higher the fidelity to correct error, as more rules can be encoded to describe the correct state of a region based on more detectable environmental variables.*

as an information backup of the original (Fig 1(i)). This is a system of trust and while imperfect, allows for the correct repair decision to be made the majority of the time. When the repair decision is incorrect, it might result in mutation: the introduction of an incorrect but dictionary legal signal element into the DNA signal. Damage can only be repaired in the context of a signal in which the damage is recognised as a dictionary illegal element. As mutation creates a legal character within the context of the signal representing the immediate DNA environment, there is not enough information on the original state of a dictionary legal signal to allow for entropy-neutral reconstitution of state, and so we can say that any dictionary legal error such as mutation is logically irreversible in the immediate context of repair enzymes. To repair this error in an entropy-neutral manner, the signal containing the error must be assessed in a higher syntax of which the local signal is but an element. We can say that the information scope must be broadened for repair.

DNA methylation sits outside the phosphate backbone and thus outside the system of trust which allows for limited local scope repair of DNA damage, and it has exactly two dictionary legal characters: fully methylated or unmethylated on both strands. Assuming no other information, this results in a situation where if one methylation is removed/added and a hemimethylated state is created, there is no logical way within the scope of a single repair enzyme to deduce which of the two legal characters the damage state originated from. The information of the original state is destroyed in the local scope, that which contains information limited to the methylation groups and immediately surrounding base pairs. We can therefore say that methylation damage is universally logically irreversible (as outlined in (**?** 17; 18)) within the local scope of repair enzymes, with all the consequences of such a trait, namely obligate entropy increase upon repair (17; 19; 20). Put simply, all hemi-state DNA methylation created by damage is the equivalent to mismatched DNA bases with intact backbones and all epigenetic damage is consequently equivalent to mutation. We can say from this that epigenetic damage repair decision-making is a recognisable but not decidable language.

Repair information is in the wrong place

In any situation of repair, the reconstitution of damage is limited by the amount of information available to the repairer. We can think of this as the scope of information that the repairer has access to. Any repair algorithm will sit within nested scopes of repair information: an enzyme might only be "aware" of the information in the immediate region of DNA it contacts; it has no access to information encoded in some distal section of DNA, or another cell, or another city. The super-entity of control represented by the system expressing and targeting that enzyme might well have access to a broader scope of information with which to target repairs. The caveat is that decisions about repair and consequently accurate repair can only occur within the scope of the information required and this may not be the scope in which the information exists. For

example, the enzyme running along DNA has more up-to-date information about the current damage-state of the piece of DNA it sits upon than does the system that sent it to fix that damage. It does not always follow that the information scope of a subunit is a subset of that of a system with a broader scope. Information loss can result in logically irreversible damage within one scope and that same damage can be logically reversible within another. The question is: which scope has access to the information necessary to detect the information loss and which scope has access to the information necessary to repair the information loss? In non-mutational DNA damage, both of these sets of required information can exist within the same scope: that of the repair enzymes. In both DNA mutation and epigenetic mutation the information encoding the state of the ontological purpose of the governed system is exclusively found within systems that have access to information on ontological outcome. The information on the identity of the specific loci (base or CpG) under interrogation is limited to the repair enzyme while the repair enzyme remains at the locus of damage. This information is temporarily segregated from that of ontological outcome: the repair enzyme will have moved on and discarded information delineating location by the time the system is observably diminished in efficacy towards survival. It is therefore impossible to provide repair enzymes with the information necessary to correctly repair specific instances of mutation after the mutation has occurred, as well as any damage that could arise from more than one dictionary legal character. The information necessary for repair is not locally available at the point repair is locally possible.

Ageing is the consequence of repair fidelity limitations

As there can be no mirroring backup to the epigenetic state and any algorithmic backup is limited below perfect fidelity, the logically irreversible information loss accrued within the epigenetic state will remain unrepairable in systems that have a complexity high enough that the information necessary for repair demands lossy compression. Damage will only be repaired up to the fidelity allowed for by the compression of repair. The only way to create logical reversibility in the systems and to reduce entropy is to increase the scope of the system until logical reversibility is possible.

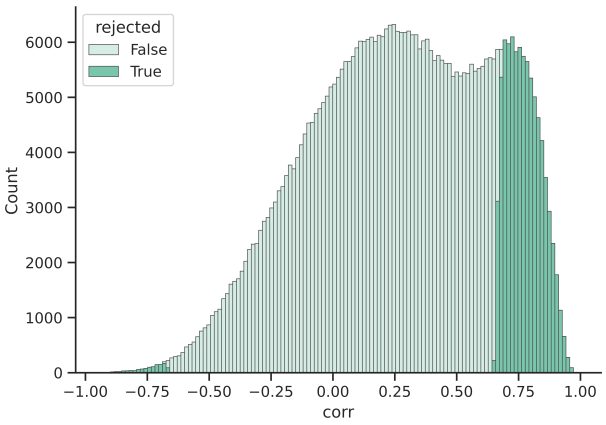
When the amount of information necessary to create a system exceeds that which is beneficially storable in DNA, lossy compression will begin to be used as information is encoded in cellular context. This is the fidelity boundary: the point beyond which perfect fidelity is impossible. By storing information in the state of the local environment, systems can minimise the need to explicitly code functionality in DNA while retaining the information for approximate functionality, but with the consequence that they become logically irreversible as the low fidelity by which they are encoded results in multiple possible original states for the current system state. At this point, repair can occur but only in an entropic manner with a degree of error.

This fidelity boundary is never reached in simple systems but when systems expand in scope to allow for logical reversibility, they increase the information necessary for repair, approaching or crossing the fidelity boundary. If any system that influences the information necessary for its own repair is complex enough to demand definition past the fidelity boundary it will imperfectly repair itself when damage occurs, generating a feedback loop as it progressively repairs itself with decreasing fidelity. We suggest that simple systems are logically reversible below the fidelity boundary and complex systems influencing their own repair are not, inevitably becoming dysfunctional unless they are so valuable for organism survival that selection encodes the entire system within DNA. Only through the construction of a true logically reversible repair scope can the inevitably accruing system corruption be fully reversed.

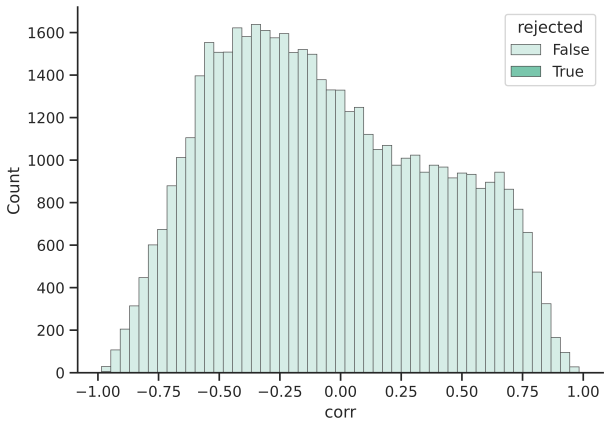
When the information required for logical reversibility exceeds that storable in the immediate context of the cell, the scope must be extended to allow for repair. Logical reversibility is then only achieved when the scope expands to include a known originator state, i.e. a stem cell. At this point, contextual algorithms with imperfect fidelity reconstitute the information of the cell (differentiation). In essence, the cell abandons its current state and returns to a point of known logical reversibility. Stem cells represent a type of cell that can be defined independently of context and thus in an informatically efficient manner. It is a simple, singular set of rules to encode, cheap and robust due to the lack of need to handle multiple definitions consequent to context. As epigenetic damage creates complexity not just in individual cells but in tissues and organs, the information defining the use of stem cells to reconstitute damage becoming itself progressively corrupted as tissue composition changes. This means that the scope that allows for logically reversible repair must be extended further back into epigenetic basality and more and more cells and eventually tissue discarded to allow for this. Eventually, this will reach such a point that childbirth is the only solution available to the organism (discarding the entire body save for a single primordial stem cell, the logical reversal of the entire organism).

Low fidelity creates error feedback

As epigenetic signaling fails, the systems governed by that signal will make incorrect decisions, resulting in a feedback cycle in which the epigenetic fidelity governing epigenetic fidelity fails, resulting in a recursive loss of epigenetic control as well as deregulation of all systems in which logical reversibility is impossible in the scope of repair. The deregulation of all cellular systems governed by epigenetic control is what, we suggest, gives rise to the



(i) *CpG SD with age correlated to age (human blood)*



(ii) *CpG SD with age correlated to age (human brain)*

Figure 3: *i. CpG SD correlation to age in GSE87571 human whole blood, using the SD to age dataset described in methods. Fluke correlation would be expected to be a normal distribution centered on zero. The peak at 0.75 represents a large population of CpG loci that increase in SD between samples with age. ii. CpG SD correlation to age in GSE41826 human brain tissue, using the SD to age dataset described in methods. As with GSE87571, we see a peak of correlation at 0.75, but a larger peak of negative correlation. We speculate that this peak represents genes switched on or off with age.*

phenotype of age.

Our theory suggests that ageing is itself the inevitable consequence of the impossibility of signal fidelity due to the specific dynamic of epigenetics being a single system in which it is impossible to design trust (through a mirrored backup) or an algorithm with perfect signal fidelity in systems where complexity is high enough that logical reversibility is impossible without crossing the fidelity boundary.

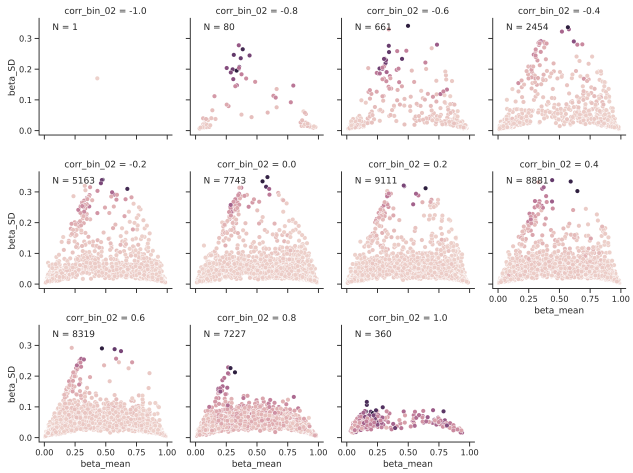
2 Results

Methylation variance with age

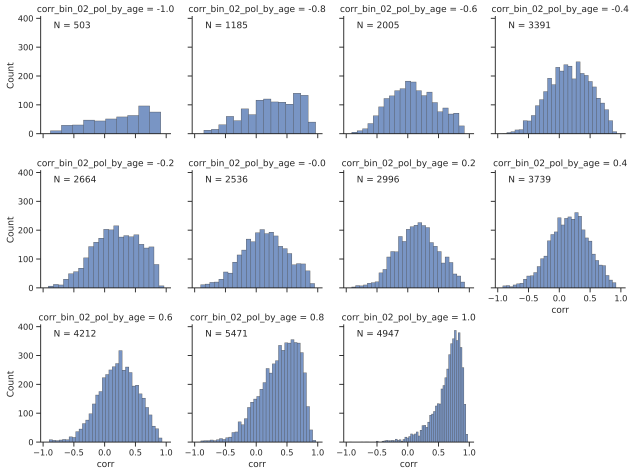
The principles outlined above suggest that there is an inevitable accumulation of epigenetic damage with age, driving the structure of epigenetic signals into randomness. One measure of this dynamic is the progressing disparity between an individual's DNA methylation loci with age. We obtained methylation data from preexisting datasets (outlined in Methods) expressed in beta values that represent the ratio of methylated to unmethylated measurements within samples for individual probe loci. For each CpG, we binned samples into age groups spanning five years and for each group obtained the standard deviation (SD) of the beta values within the group. We then performed a Pearson's R correlation between the age-binned SD and age. Results are summarised in Fig 3(i) and 3(ii), and in supplementary table 1.1 (human blood) and supplementary table 1.2 (human brain). We used Benjamini-Hochberg correction to account for multiple testing, but most forms of multiple testing are heavily biased to extremely strong correlation, and in any analysis of stochastic noise the understanding of what represents 'fluke' correlation can be observed through the expectation that these will be represented by a normal distribution centered on 0 correlation. In all tissues, we observe a non-normal distribution of correlation. In long-lived mammals, we observe a conserved peak within GSE87571 (human blood), GSE41826 (human brain), and GSE184223 (zebra blood) around r=0.75. In human blood, we observed another peak at approx. r=0.12 and in human brain we observe a peak at r=-0.3. In GSE120137 (mice tissues) we observe a single distribution centered on r=0.75 in blood, an even more extreme distribution centered on r=0.9 in muscle, a three-pole distribution with peaks at r=-1.00, r=0.00, and r=1.00 in adipose and kidney, with lung and liver also having peaks at r=-1.00 and r=1.00 but with the peak at 0.00 unpronounced, probably hidden by merging with the tails of the polar peaks. The more pronounced polarisation of r values in this dataset was likely consequent to there only being three recorded age groups, 2, 10, and 20 months. The general observation is that there are three approximate classes of loci: those that correlated negatively with age, those that correlate positively, and those representing 'fluke' correlation, centered on r=0.00.

Methylation polarisation with age

We theorised that those CpG with strong positive correlation between SD and age represented a class of noise-retaining loci. We expected a peak at r=0.00 to represent fluke correlation, and we next set out to characterise the nature of the negative correlations. We suggest that as samples are getting older, there is a tendency to switch on or off genes as a mechanism to control



(i) *Beta distributions by correlation*



(ii) *Polarisation of CpGs by correlation*

Figure 4: *i. CpG beta value vs CpG SD by CpG SD correlation to age in GSE87571 human whole blood. As beta SD correlation to age increases, the loci with higher SD decrease in representation. ii. Correlation of CpG SD to age faceted by centralisation correlation to age in GSE87571 Human whole blood. CpG loci in which SD correlated heavily to age are those in which beta centralisation increases with age.*

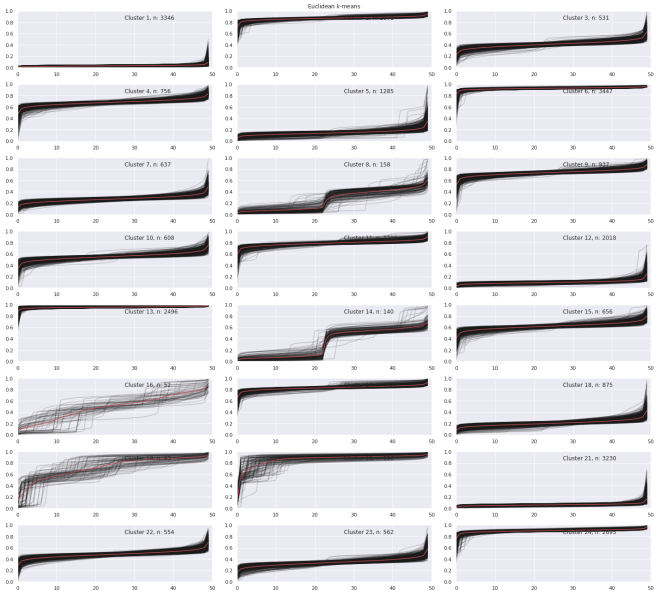
noise. We theorised that these genes would therefore tend towards polar beta values in their regulating methylation (either fully methylated or unmethylated). To explore this, we segregated loci SD correlation to age by the polarisation of their mean beta value with age. This demonstrated that in all datasets (GSE87571 (human blood), GSE41826 (human brain), GSE184223 (zebra blood), GSE120137 (mouse tissues)) the distribution of negative correlation is heavily skewed towards polarising loci, and that across all tissue datasets the peak at r=0.75 is retained, segregated to those CpG in which centralisation of beta value increases with age. This suggests that those loci with that decrease in variance with age indeed represent those genes that are increasingly regulated in response/consequent to age, and thus a different class of loci to those that, free of regulation, drift into variance.

Methylation topology is an activation function

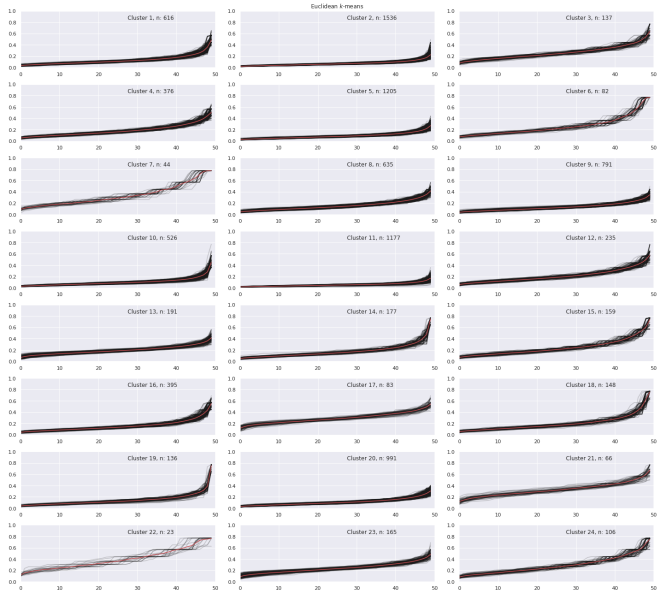
We hypothesised that if different methylation regions represented different classes of epigenetic control resulting from the need for discrimination in the amount of noise gene functionality was exposed to, evolution's naturally conserving effect would unify these classes into a few different regulatory activation functions governed by a simple set of arguments. Were this the case, we would expect to observe conservation between the way specific loci were themselves regulated and therefore regulated the underlying gene. To visualise this, we performed Euclidean k-means clustering on sorted preparation of all datasets (Fig 5(i)). In GSE87571 (human blood), GSE41826 (human brain), and GSE184223 (zebra blood) we observed a definite grouping in locus topology, indicating that there are a few archetypical "activation functions" to which all DNA methylation belongs, and by which all methylation is regulated. It appears that DNA methylation is controlled by two types of functions, a linear function and a step function, which themselves act as components for a small range of combination activation functions. We classify these overall activation functions as linear functions, single-step functions, multistep functions, and "ragged" functions (functions containing regions with numerous fractional subpopulations independent of the majority ontological state). The topology of GSE120137 (mouse tissues) differs from that of longer-lived mammals (Fig 5(ii)). Both linear and step functions are observable, but the banding effect seen in human methylation data is absent.

Methylation topology 'tailing'

Examining the topology of the clusters and individual CpG loci highlights that in most cases there is a majority position output by the function controlling methylation for each individual CpG locus. We assume that these are the positions for the ontological purpose of system function. In a large number of loci, there is also a 'tail' region of rapidly increasing or decreasing methylation approaching the nearest occupancy absolute (beta 0.00/1.00). These tails have the distinguishing feature of being an approximately consistent fraction of the total population size, but the difference between the tip of the tail and the mean beta value can vary substantially. We theorise that these tails represent a failure of control, in which a methylation area that is ideally at a given level of methylation loses its ability to regulate itself, resulting in mean betas that differ greatly from the ontological target. Mouse topologies increase at a comparatively steep and smooth linear rate



(i) *Methylation activation functions (human blood)*



(ii) *Methylation activation functions (mouse kidney)*

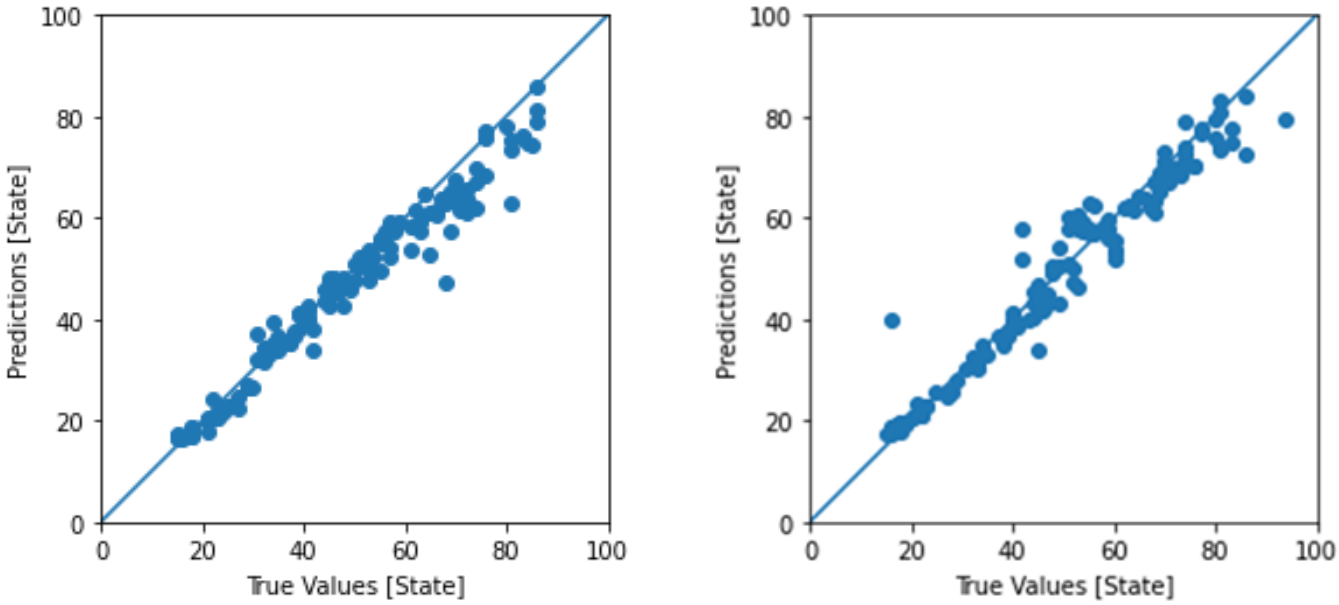
Figure 5: *i. Topological activation functions of DNA methylation in GSE87571 human whole blood. Sorted dataset clustered using Euclidean k-means. ii. Topological activation functions of DNA methylation in GSE120137 mouse kidney. Sorted dataset clustered using Euclidean k-means. (x resampled to 50)*

with very little evidence of the "tailing effect" seen in the loci of longer-lived mammals (Fig 5(ii)), which suggests that there is comparatively less 'failure of regulation' because there is less regulation in the first place, mice not selecting as strongly for epigenetic fidelity as mammals more exposed to epigenetic mutation through lifespan and lack of predation. It is notable that these tails do not correlate to age in all loci. We believe that these may in some cases represent a loss of control in disease states and/or systems that lack the ability to create positive feedback into further epigenetic deregulation.

Methylation topology 'walling'

To gain a deeper understanding of the observed dynamics, we created a state machine that modeled regions of CpGs under variable methylation pressure, which we defined as the ratio between the overall frequency of methylation addition against that of methylation removal. Rather than defining CpG islands (CGI) holistically, we modeled methylation regions as being 'soft edged', representing the normal distribution of randomly walking methylation enzymes and other local effects. Independently of these local effects, we modeled uniform static pressures such as stochastic loss of methylation and other single entity dependent effects.

Within this model, we observed the existence of two universal effects within any given region of methylation. First, there is a methylation boundary effect determined by current region occupancy. The current occupancy level of a methylation region creates a proportional diminishment in the ratio of methylation pressure to methylation state change, by which can be concluded that the efficacy of methylation pressure has an inverse relationship to the direction the pressure is traveling. Taking observed global peaks of methylation states as parameters (approx. 0.125/0.925 occupancy in global methylation (21; 22) we estimated that the methylation pressure necessary to move a peak from 0.125 to 0.925 is 180/1 positive to negative pressure, i.e. there needs to be a 180 times more addition than removal of methylation to result in an average methylation occupancy of 0.925. This means that as average methylation approaches the occupancy walls of 0 and 1 beta, greater pressures must be applied to the regions, and any deviation from the mean beta value in loci close to the walls represents a far more extreme failure of control than deviation further away from the walls. Secondly, the induction of noise pushes the equilibrium towards 0.5 occupancy ratio overall, but only in the biologically unlikely scenario that noise is uniform across pressure components. Stochastic loss of methylation, for example, seems unlikely to become noisy consequent to deregulation because it is by definition unregulated. When noise is added to individual pressure components, those components have their overall pressure coefficient diminished towards zero, meaning that we should not necessarily expect noise from the deregulation of epigenetic maintenance to necessarily approach homogenous occupancy.



(i) *Classification of age by deviation*

(ii) *Classifcation of age by ranked deviation*

Figure 6: *Classifiers trained on a sliding window dataset (i.) and a ranked SD dataset (ii.) of GSE87571 human whole blood DNA methylation beta values(window size = 5, as described in methods). Chronological age (y) is measured against vs predicted age (x), with perfect prediction being represented by the samples in which the predicted age matches the chronological age.*

Neural Networks can predict age from noise

We trained a Keras classifier with a range of parameters (supplementary table 2.1) on both the blood and brain 'SD by age', 'sliding window SD by age', and 'ranked SD by age' datasets. Features were selected by random sampling and the results represent the mean of a Kfold split (f=5). Results are summarised in supplementary table 2.2. We observe that age can be predicted with great accuracy from individual loci beta SD within age groups, illustrating neural networks' capability to predict based entirely on the amount of variance in a CpG site, rather than information on actual methylation levels (Fig 6(i)).

To remove all possibility that the loci SD provided to the network was allowing it to estimate beta values, we used a dataset where locus SD values are replaced with a rank based on size, preventing the network from having any understanding of the extent of the SD and leaving only the information about how sample variance ranked against other samples at that locus (Fig 6(ii)). As a result, the network lost only a small amount of predictive power, with a mean absolute error of 2.573 (R2 0.962), which demonstrates that prediction is possible solely based on the knowledge of relative deviation within CpG sites.

Control is lost in the regulation of regulation

To improve the classification of loci in which SD most correlates to age, we reasoned that 'fluke' correlation was more likely to be represented in individual CpG loci than entire CGI which should be more consistent in their ontological effect. To this end, we grouped loci by their CGI and calculated the SD of the correlations within these CGI.

We subset all used human datasets to only include CpG from CGI with a correlation SD below 0.15, and further subset to contain only CpG in which SD to age correlation and centralisation were above 0.8. We then used methylGSA (23) to perform gene set analysis on unique genes associated with these loci and compared them against both those loci that fall outside this criteria and ten sets of randomly sampled of loci(n=100000).

Within the human genome datasets this preparation resulted in the enrichment of sequence-specific DNA binding (GO:0043565), RNA polymerase II transcription regulatory region sequence-specific DNA binding (GO:0000977), cis-regulatory region sequence-specific DNA binding (GO:0000987) and other terms that relate to transcriptional regulation (summarised in supplementary table 3.1 (GSE87571 human blood) and 3.2 (GSE41826 human brain)), traits absent from the randomly sampled sets and the low polarisation set. It is interesting that those CGI in which SD correlates to age are those that are involved with the regulation of promoters and enhancers, i.e. the regulation of regulation.

3 Discussion

The role of epigenetic changes in ageing has been a major research focus, in particular since the discovery of epigenetic clocks. In this work we set out to break down the nature of epigenetic damage and characterise biological ageing as a failure of repair fidelity. To do this, we began by showing that chronological age correlates to the progressive deviation within certain classes of CpG loci. It seems that there are three variables that control the distribution of SD correlation to age: those values representing fluke correlation to cellular regulation centered near 0, those representing genes that are being regulated increasingly as age increases (senescence, DNA repair, stress response, etc) and those peaks describing epigenetic systems becoming deregulated with age (Fig 3(i)). The latter approach random methylation and so are detectable by their tendency to centralise their average beta. We suggest that epigenetic clocks are measuring both of the latter classes of CpG and that the answer to the question 'Are epigenetic clocks measuring cause or effect?' (5) is 'both', depending on the CpGs used. Genes that change in response to age represent the effect of age and the epigenetic stochastic noise represents biological age itself.

We also suggest that there is evidence that there are 'activation functions' of regional epigenetic control that are observable through the limited range of outputs that define average methylation across populations (Fig 5(i)). These activation functions are defined through control of local methylation pressure, upon which will be applied mechanisms that we suspect to be a handful of evolutionarily conserved behaviours repeatedly applied in combination to produce a limited range of robust behaviours.

Within these activation functions, we suggest that deviation from the ontologically intended output of the activation function represents epigenetic damage, eventually resulting in deregulation of the governed gene expression. It seems likely that this is what is represented by the topology 'tails', samples in which loci control has either become aberrant (or possibly some cases where the loci were measured during a transition state).

Our analysis shows that as beta SD correlation to age increases, mean beta SD decreases (Fig 4(i)). Combined with our conclusion above, this paints a profile of low SD unwalled loci as being the most correlated with age. We suggest that activation function deregulation is more extreme in situations where noise can push deregulation in both directions, as opposed to only away from a wall. These would be the most difficult loci in which to maintain precise control, while also being loci in which precise control is required, as evidenced by their overall minimal variance. We suggest that there is a selection pressure towards regulating by polarisation, minimising variance by creating pressure opposed by occupancy effects. Those loci that do not have outputs aimed at placing them close to an extreme must have a selection pressure superseding these benefits. There seems to be a reason that some loci are less regulated, and thus more prone to retaining error, and these loci are those in which SD correlates to age.

We can demonstrate that these loci are separate in class from genes that are regulated in response to age (Fig 4(ii)) and further demonstrate through the ranked variance clock that age can be predicted exclusively from the knowledge of this deregulation (Fig 6(ii)). This fits exactly with our initial theory: epigenetic damage is inevitable due to the impossibility of mirrored backup and the bounding that limits any algorithmic repair to imperfect fidelity.

We propose that this epigenetic damage would result in a feedback cycle, in which deregulation would lead to further deregulation through the disruption of maintenance and repair of the epigenetic regions, and to the phenotype of age through the general deregulation of cellular systems. This would fit the profile of ageing as a robust, gradual process, with slow, reliable progress made as deregulation accumulates, accelerating toward network failure as the feedback cycle picks up pace. We can see in the gene ontology results that in all organisms and tissues, those genes regulated by the loci in which deregulation correlates to age are genes governing promotors and enhancers.

We suggest that this is because promotors and enhancers have a unique feature that precludes polarising their regional control for regulation: they need to regularly reconfigure the local methylation state consequent to the current state of transcription. We suggest this makes them tautologically defined, in that the definition of the epigenetic signal of a promotor/enhancer modulator relies in part on its own current state (such that any damage results in damage to any rule from which the signal could be corrected), and thus representing a class of loci in which epigenetic regional control cannot be correctly defined once epigenetic damage has occurred. We suggest damage accrues in these regions and the global deregulation of transcription that occurs consequent to this gives rise to the general phenotype of age.

Our work demonstrates that the cell cannot perfectly reconstitute epigenetic damage whenever it occurs, and must instead pick and choose which systems to hold to high fidelity and which to allow low fidelity. In some cases, we argue, low fidelity is forced either way. There can be no perfect fidelity backup of all epigenetic information stored in the cell.

Our work is broadly consistent with the information loss theory of ageing (12; 11), but we suggest that the cell doesn't have a problem with information loss: information is lost all the time. What the cell has an issue with is fidelity of repair, and this ultimately boils down to a limitation of data storage and trust sourcing. We also argue that it is not merely systemic damage: epigenetic damage causes 'definition damage', as the tautological

nature of epigenetic regulation and epigenetic definition causes certain systems to treat accepted damage as 'the new normal'.

We would also argue against the idea that any specific cellular subsystem explains the general dynamic of age, rather suggesting that the frequency of epigenetic events is likely to be the main influence of the rate of epigenetic damage and it is a core dynamic of epigenetic control that information loss must occur in the regulation of any adaptive system. The nature of the systems themselves is irrelevant to the frequency of epigenetic modification they demand and the fact that such modification must in these systems be tautologically defined. We argue that DNA repair and other correlated behaviours are simply representatives of this class of epigenetic regulation and those models that recognise the link between age and particular systems such as DNA repair (12; 11) are measuring the rate at which epigenetic modification within these systems leads to epigenetic damage. It has been demonstrated that there is a general stochastic loss of methylation over time (24) and that hemimethylated states can spontaneously occur through methyl group drop-off or enzymatic error (25), and we suggest that this might provide a base rate at which biological ageing progresses.

This theory, therefore, provides a full line of reasoning from the logical necessity of epigenetic damage through to the phenotype of age. The outlined theory is not limited to methylation-based damage: the duplication of histone markers and damage within that system shares the exact same issue as methylation, as will any outer layer system of information fidelity. In an effort to reduce the information necessary to maintain a system in which damage is logically reversible, we propose mammals use childbirth to shed complexity, reducing the information necessary to a single cell. In primordial cells, stem cells and the simple cellular systems of organisms without epigenetic regulation, damage is reversible through the encoding of state in algorithms that can define the correct informatic state within the genetic code. This is possible due to the ability to prescribe what a cell should be independently of its environment, and results in logical reversibility. We suggest that logical reversibility is a necessary objective for any organism existing in a noisy channel.

4 Methods

Datasets were handled in Python (v3.8.10) using Pandas (v1.3.5) and visualised with a combination of Matplotlib (v3.1.3) and Seaborn (v0.11.2). Datasets were obtained by direct download from the NCBI GEO repository (https://www.ncbi.nlm.nih.gov/) or where unavailable accessed through the Python package GEOparse (v2.0.3). Datasets were acquired with the most recent version as of 01/09/2022. GSE87571 is GPL13534 (450k array) beadchip data from human whole blood. Processed data was used as described in the original paper (26). In brief, the dataset was 728 samples from 421 individuals with an age range of 14 to 94 years old. GSE41826 is GPL13534 (450k array) beadchip data from human brain tissue. Processed data was used as described in the original paper (27). In brief, the dataset was 145 samples from 58 individuals with an age range of 13 to 79 years old. Half the samples are from healthy controls and half from subjects with major depression. GSE120137 is GPL21103 (Illumina HiSeq) data from mouse liver, lung, adipose, blood, kidney and muscle tissue. Processed data was used as described in the original paper (28). GSE184223 is GPL28271 (Illumina HorvathMammalian-MethylChip40) data from zebra blood and biopsy (only blood used). Processed data was used as described in the original paper (29).

These datasets were used as a base to create additional datasets under the following conditions:

SD by age datasets

Samples were binned into groups covering five years of age (e.g. 60-65 years old). Within each bin, the standard deviation was taken for the beta values within each individual CpG loci, resulting in a table of age group SD by loci.

Sliding window SD by age datasets

The sliding window SD by age dataset was created by sorting samples by age and running a five-row sliding window on each CpG taking the SD for each window step. Sample order was then re-randomised.

Ranked SD by age datasets

The ranked SD dataset was created by ranking the sliding window dataset within individual loci.

Polarisation datasets

The polarisation datasets were created with the following transformation on every value (x):

$$0.5 - |X - 0.5|$$

Statistical Tests

Statistical tests were performed using scikit-learn (v1.0.2) and Benjamini-Hochberg multiple correction was handled using the 'multipletests' function from statsmodels (v0.12.2). Beta plots were resampled using 'TimeSeriesResampler' and clustering was performed using 'TimeSeriesKMeans' from the package tslearn (v0.5.3.2). A random seed of 0 was used for all clusterings. Pearson's R correlation was performed with the pandas 'corrwith' function, feature ranking and sliding windows were performed using the pandas 'rank' and 'rolling'

functions respectively, and feature sorting was performed with the Python built-in 'sorted' function.

Deep Learning and State Machine

Deep learning was performed using Tensorflow (v.2.11.0) Keras. The methylation state machine was coded in Python (v3.8.10).

Gene Ontology

GSEA was performed using the R package methylGSA (23) (v.1.16.0, gene set size minimum = 100, maximum = 500, method RRA(GSEA), using the possible genes associated to GPL13534 as background.

Data Availability

Correlation results are made available in supplementary tables. State machine code will be made available upon publication.

Correspondence to: Thomas Duffield (email: thomas.duffield.correspondence@gmail.com) or João Pedro de Magalhães (email: jp@senescence.info)

Work in our lab is supported by grants from the Wellcome Trust (208375/Z/17/Z), Longevity Impetus Grants, LongeCity and the Biotechnology and Biological Sciences Research Council (BB/R014949/1 and BB/V010123/1).

Conflicts of interest: JPM is CSO of YouthBio Therapeutics, an advisor/consultant for the Longevity Vision Fund and NOVOS, and the founder of Magellan Science Ltd, a company providing consulting services in longevity science.

References

[1] L. Partridge, J. Deelen, and P. E. Slagboom, "Facing up to the global challenges of ageing," *Nature*, vol. 561, pp. 45–56, Sept. 2018. Number: 7721 Publisher: Nature Publishing Group.

[2] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "Hallmarks of aging: An expanding universe," *Cell*, vol. 186, pp. 243–278, Jan. 2023. Publisher: Elsevier.

[3] D. Gems and J. P. de Magalhães, "The hoverfly and the wasp: A critique of the hallmarks of aging as a paradigm," *Ageing Research Reviews*, vol. 70, p. 101407, Sept. 2021.

[4] K. Seale, S. Horvath, A. Teschendorff, N. Eynon, and S. Voisin, "Making sense of the ageing methylome," *Nature Reviews Genetics*, vol. 23, pp. 585–605, Oct. 2022. Number: 10 Publisher: Nature Publishing Group.

[5] S. Horvath and K. Raj, "DNA methylation-based biomarkers and the epigenetic clock theory of ageing," *Nature Reviews Genetics*, vol. 19, pp. 371–384, June 2018. Number: 6 Publisher: Nature Publishing Group.

[6] B. A. Benayoun, E. A. Pollina, and A. Brunet, "Epigenetic regulation of ageing: linking environmental inputs to genomic stability," *Nature Reviews Molecular Cell Biology*, vol. 16, pp. 593–610, Oct. 2015. Number: 10 Publisher: Nature Publishing Group.

[7] S. Horvath, "DNA methylation age of human tissues and cell types," *Genome Biology*, vol. 14, p. 3156, Dec. 2013.

[8] R. E. Marioni, S. Shah, A. F. McRae, B. H. Chen, E. Colicino, S. E. Harris, J. Gibson, A. K. Henders, P. Redmond, S. R. Cox, A. Pattie, J. Corley, L. Murphy, N. G. Martin, G. W. Montgomery, A. P. Feinberg, M. D. Fallin, M. L. Multhaup, A. E. Jaffe, R. Joehanes, J. Schwartz, A. C. Just, K. L. Lunetta, J. M. Murabito, J. M. Starr, S. Horvath, A. A. Baccarelli, D. Levy, P. M. Visscher, N. R. Wray, and I. J. Deary, "DNA methylation age of blood predicts all-cause mortality in later life," *Genome Biology*, vol. 16, p. 25, Jan. 2015.

[9] K. Raj and S. Horvath, "Current perspectives on the cellular and molecular features of epigenetic ageing," *Experimental Biology and Medicine*, vol. 245, pp. 1532–1542, Nov. 2020. Publisher: SAGE Publications.

[10] A. A. Johnson, K. Akman, S. R. Calimport, D. Wuttke, A. Stolzing, and J. P. de Magalhães, "The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease," *Rejuvenation Research*, vol. 15, pp. 483–494, Oct. 2012. Publisher: Mary Ann Liebert, Inc., publishers.

[11] J.-H. Yang, M. Hayano, P. T. Griffin, J. A. Amorim, M. S. Bonkowski, J. K. Apostolides, E. L. Salfati, M. Blanchette, E. M. Munding, M. Bhakta, Y. C. Chew, W. Guo, X. Yang, S. Maybury-Lewis, X. Tian, J. M. Ross, G. Coppotelli, M. V. Meer, R. Rogers-Hammond, D. L. Vera, Y. R. Lu, J. W. Pippin, M. L. Creswell, Z. Dou, C. Xu, S. J. Mitchell, A. Das, B. L. O'Connell, S. Thakur, A. E. Kane, Q. Su, Y. Mohri, E. K. Nishimura, L. Schaevitz, N. Garg, A.-M. Balta, M. A. Rego, M. Gregory-Ksander, T. C. Jakobs, L. Zhong, H. Wakimoto, J. E. Andari, D. Grimm, R. Mostoslavsky, A. J. Wagers, K. Tsubota, S. J. Bonasera, C. M. Palmeira, J. G. Seidman, C. E. Seidman, N. S. Wolf, J. A. Kreiling, J. M. Sedivy, G. F. Murphy, R. E. Green, B. A. Garcia, S. L. Berger, P. Oberdoerffer, S. J. Shankland, V. N. Gladyshev, B. R. Ksander, A. R. Pfenning, L. A. Rajman, and D. A. Sinclair, "Loss of epigenetic information as a cause of mammalian aging," *Cell*,

vol. 186, pp. 305–326.e27, Jan. 2023. Publisher: Elsevier.

[12] D. Sinclair, *Lifespan: The Revolutionary Science of Why We Age and Why We Don't Have To*. London: Thorsons, an imprint of HarperCollinsPublishers, uk edition ed., 2019.

[13] J. P. de Magalhães, "Ageing as a software design flaw," *Genome Biology*, vol. 24, p. 51, Mar. 2023.

[14] R. Cameron, "Infinite Regress Arguments," July 2018.

[15] G. Bonino, "Bradley's Regress: Relations, Exemplification, Unity," *Axiomathes*, vol. 23, pp. 189–200, June 2013.

[16] L. A. Fitriya, T. W. Purboyo, and A. L. Prasasti, "A Review of Data Compression Techniques," *International Journal of Applied Engineering Research*, vol. 12, no. 19, pp. 8956–8963, 2017.

[17] R. Landauer, "Dissipation and noise immunity in computation and communication," *Nature*, vol. 335, pp. 779–784, Oct. 1988. Number: 6193 Publisher: Nature Publishing Group.

[18] B. Hayes, "Reverse Engineering," *American Scientist*, vol. 94, no. 2, p. 107, 2006.

[19] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, July 1961. Conference Name: IBM Journal of Research and Development.

[20] R. Landauer, "Minimal Energy Requirements in Communication," *Science*, vol. 272, pp. 1914–1918, June 1996. Publisher: American Association for the Advancement of Science.

[21] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, vol. 15, p. R37, Feb. 2014.

[22] V. Kukushkina, V. Modhukur, M. Suhorutšenko, M. Peters, R. Mägi, N. Rahmioglu, A. Velthut-Meikas, S. Altmäe, F. J. Esteban, J. Vilo, K. Zondervan, A. Salumets, and T. Laisk-Podar, "DNA methylation changes in endometrium and correlation with gene expression during the transition from pre-receptive to receptive phase," *Scientific Reports*, vol. 7, p. 3916, June 2017. Number: 1 Publisher: Nature Publishing Group.

[23] X. Ren and P. F. Kuan, "methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing," *Bioinformatics*, vol. 35, pp. 1958–1959, June 2019.

[24] T. Chen, Y. Ueda, J. E. Dodge, Z. Wang, and E. Li, "Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b," *Molecular and Cellular Biology*, vol. 23, pp. 5594–5605, Aug. 2003. Publisher: Taylor & Francis _eprint: https://doi.org/10.1128/MCB.23.16.5594-5605.2003.

[25] J. Arand, D. Spieler, T. Karius, M. R. Branco, D. Meilinger, A. Meissner, T. Jenuwein, G. Xu, H. Leonhardt, V. Wolf, and J. Walter, "In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases," *PLOS Genetics*, vol. 8, p. e1002750, June 2012. Publisher: Public Library of Science.

[26] A. Johansson, S. Enroth, and U. Gyllensten, "Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan," *PloS One*, vol. 8, no. 6, p. e67378, 2013.

[27] J. Guintivano, M. J. Aryee, and Z. A. Kaminsky, "A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression," *Epigenetics*, vol. 8, pp. 290–302, Mar. 2013.

[28] M. J. Thompson, K. Chwiałkowska, L. Rubbi, A. J. Lusis, R. C. Davis, A. Srivastava, R. Korstanje, G. A. Churchill, S. Horvath, and M. Pellegrini, "A multi-tissue full lifespan epigenetic clock for mice," *Aging*, vol. 10, pp. 2832–2854, Oct. 2018.

[29] S. Horvath, A. Haghani, S. Peng, E. N. Hales, J. A. Zoller, K. Raj, B. Larison, T. R. Robeck, J. L. Petersen, R. R. Bellone, and C. J. Finno, "DNA methylation aging and transcriptomic studies in horses," *Nature Communications*, vol. 13, p. 40, Jan. 2022.