Universal DNA methylation age across mammalian tissues

MAMMALIAN METHYLATION CONSORTIUM

Correspondence: shorvath@mednet.ucla.edu

ABSTRACT

Aging is often perceived as a degenerative process caused by random accrual of cellular damage over time. In spite of this, age can be accurately estimated by epigenetic clocks based on DNA methylation profiles from almost any tissue of the body. Since such pan-tissue epigenetic clocks have been successfully developed for several different species, it is difficult to ignore the likelihood that a defined and shared mechanism instead, underlies the aging process. To address this, we generated 10,000 methylation arrays, each profiling up to 37,000 cytosines in highly-conserved stretches of DNA, from over 59 tissue-types derived from 128 mammalian species. From these, we identified and characterized specific cytosines, whose methylation levels change with age across mammalian species. Genes associated with these cytosines are greatly enriched in mammalian developmental processes and implicated in age-associated diseases. From the methylation profiles of these age-related cytosines, we successfully constructed three highly accurate universal mammalian clocks for eutherians, and one universal clock for marsupials. The universal clocks for eutherians are similarly accurate for estimating ages (r>0.96) of any mammalian species and tissue with a single mathematical formula. Collectively, these new observations support the notion that aging is indeed evolutionarily conserved and coupled to developmental processes across all mammalian species - a notion that was long-debated without the benefit of this new and compelling evidence.

INTRODUCTION

Aging is associated with multiple cellular changes that are often tissue-specific. Cytosine methylation however, is unusual in this regard as it is strongly correlated with age across virtually all tissues. This feature can be capitalized upon to develop multivariate age-estimators (pan-tissue epigenetic clocks) that are applicable to most or all tissues of a species. This approach produced the first human pan-tissue clock that was based on 353 age-related CpGs [1]. Subsequent successes in developing similar pan-tissue clocks for other species hint at the universality of the aging process. To investigate this, we sought to i) identify and characterize cytosines whose methylation levels change with age in all mammals and ii) develop universal age-estimators that apply to all mammalian species and tissues (universal epigenetic clocks for mammals). Towards these ends, we employed a novel Infinium array (HorvathMammalMethylChip40) that profiles methylation levels of up to 37k CpGs with flanking DNA sequences that are highly-conserved across the mammalian class [2]. We obtained such profiles from almost 10,000 samples from 59 tissue types, derived from 128 mammalian species, representing 15 phylogenetic orders **(Supplementary Tables 1.1-1.5)** with age ranging from prenatal, to 139-years-old (bowhead whale). The species tested had maximum life spans from 3.8 to 211 years and adult weights from 0.004 to 100,000 kilograms.

To identify age-related CpGs, we carried out two-stage meta-analysis across species and tissues. Cytosines that become increasingly methylated with age (*i.e.*, positively correlated) were found to be more highly conserved (**Fig. 1a**). From these, we identified 665 age-related CpGs, within a threshold significance of $\alpha=10^{-200}$ across all eutherian species and tissues **(Fig. 1a, Supplementary Table 2.1)**. Cytosines cg12841266 ($P=6.2\times10^{-908}$) and cg11084334 ($P=2.0\times10^{-823}$), located in exon 2 of the *LHFPL4* gene were the most predictive across all species, having a correlation >0.8 in 24 species **(Supplementary Table 3),** of which three are shown in **Fig. 1b-d**. Another highly-correlated cytosine, cg09710440, resides in *LHFPL3* ($P= 5.1\times10^{-724}$), a paralog of *LHFPL4* **(Fig. 1a, Extended Data Fig. 1, Supplementary Table 2.1)**. As *LHFPL4* and *LHFPL3* are in human chromosomes 2 and 7 respectively,

their age-related gain of methylation is unlikely to be random. It implies instead their involvement in the aging process, even if their activities as nucleators of GABA receptors do not immediately conjure an obvious mechanism. Indeed, methylation of *LHFPL4* cg12841266 was strongly correlated with age of multiple mouse tissues in both development (r=0.58 and P=8.9x10^{-11}) and post-development stages (r=0.45 and P=2.3x10^{-76}), particularly in the brain (r=0.92 and P=6.95x10^{-8}), muscle (r=0.89 and P=7.6x10^{-7}), liver (r=0.79 and P=1.9x10^{-117}), and blood (r=0.89 and P=1.0x10^{-53}, **Extended Data Fig. 2**). Consistent with increased methylation, expression of both *LHFPL4* and *LHFPL3* declines with increasing age in numerous, albeit not all, human and mouse tissues (**Supplementary Tables 4.1-4.4**). In particular, their reduced expression is consistently observed in the brain [3,4]. Importantly, age-related methylation changes in young animals concur strongly with those observed in middle-aged or old animals, excluding the likelihood that the changes are those involved purely in the process of organismal development (**Extended Data Figs. 3 and 4**).

Meta-analysis of age-related CpGs across specific tissues

To gain a wider and deeper understanding of age-related CpGs within specific tissues across different species, we focused on 5 organs: brain (whole and cortex), blood, liver, muscle and skin. We performed EWAS meta-analysis on 851 whole brains (17 species), 391 cortices (6 species), 3552 blood (28 species), 1221 liver (9 species), 345 muscle (5 species), and 1452 skin (31 species). Consistently across all tissues, there were more CpG with positive correlations with age than negative ones (**Extended Data Fig. 1)** and most of them were located within CpG islands, which are known to become increasingly methylated with age (**Fig. 1f, Supplementary Tables 2.2-2.7)**. While many of these cytosines were either specific to individual organs or shared between several organs, 54 potential universal age-related CpGs were shared among all the five organs **(Fig. 1e, Extended Data Table 1)**. Strikingly, the overwhelming majority of the 36 genes that are proximal to these 54 CpGs encode transcription factors with homeobox domain, and are involved in developmental processes (**Extended Data Table 1**).

Functional enrichment analysis of age-related CpGs

We employed a pathway enrichment tool (GREAT hypergeometric test based on genomic regions[5]) to analyze the top 1,000 positively and 1,000 negatively correlated age-related CpGs and their proximal genes in all tissues, individually or collectively, to ascertain whether they are associated with particular biological processes or cellular pathways (**Fig. 1g**, **Supplementary Tables 5.1-5.15)**. We demonstrated that our enrichment results are not confounded by the special design of the mammalian methylation array (**Supplementary Information**, **Note 2**). From positively-correlated CpGs across all tissues, the most enriched (P=3.7x10^{-207}) Gene Ontology term was "nervous system development", which also appeared prominently in blood (P=4.7x10^{-230}), liver (P=7.6x10^{-136}), muscle (P=1.4x10^{-12}), skin (P=5.4x10^{-141}), brain (P=1.0x10^{-42}) and cortex (P=7.5x10^{-80}). Other top-scoring terms include “pattern specification” and “anatomical structure development” (**Extended Data Table 2 & Supplementary Table 5s)**. Evidently, many hypermethylated age-related CpGs in all the five organs may be proximal to development genes. At the molecular level, many of these CpGs are in positions targeted by SUZ12, which is one of the core subunits of polycomb repressive complex 2 (all tissue P=7.1x10^{-225}, blood P=3.9x10^{-259}, liver P=1.7x10^{-149}, muscle P=8.2x10^{-16}, skin P=2.6x10^{-150}, brain P=8.7x10^{-54} , and cerebral cortex P=6.1x10^{-87}); echoing previous human EWAS studies[6,7]. EED, another core subunit of PRC2, shows similarly high significant P-values, e.g. P=1.7x10^{-262} in all tissues (**Extended Data Table 2**). Strong enrichment can also be found in promoters with H3K27me3 modification. These were observed in all tissue (P=2.8x10^{-266}), blood (P=3.9x10^{-283}), liver (P=3.3x10^{-189}), muscle (P=8.7x10^{-18}), skin (P=3.3x10^{-189}), brain (P=3.3x10^{-68}), and cortex (P=5.1x10^{-116}) (**Extended Data Table 2**). These results reinforce the association between development and aging.

This may appear counterintuitive but finds support from the fact that mice with compromised development following ablation of growth hormone receptors (GHRKO), exhibit significant slowing

down of their aging process [8]. We demonstrated that the universal epigenetic clocks are slowed in cortex, liver, and kidneys from GHRKO mice (**Extended Data Fig. 4)**.

Interestingly, although there were 3,617 enrichments of hypermethylated age-related CpGs across all tissues, only 12 were found for hypomethylated ones. The apparent scarcity of the latter is influenced by enrichment asymmetry that is particularly strong in skin, blood, and liver (**Supplementary Table 5.1**). However, this is not the case for the brain, cerebellum, cortex, and muscle, where there was instead greater enrichment of hypomethylated age-related cytosines; a trend that seemingly parallels the rate of tissue turn-over. The cytosines that were negatively associated with age in brain and cortex, but not skin, blood, and liver, are enriched in the circadian rhythm pathway ($P \geq 9.0 \times 10^{-18}$, **Supplementary Tables 5.5, 5.7, Extended Data Table 2**), indicating that besides commonly shared processes of development, which is universally implicated in aging of all tissues, organ-specific ones are also clearly in operation.

Another relevant observation is the enrichment of negative age-related cytosines in an up-regulated gene set in Alzheimer's disease. This was observed in the whole brain ($P=2.1 \times 10^{-30}$, **Extended Data Table 2**), the cortex ($P=5.9 \times 10^{-22}$), and in muscle tissue ($P=2.5 \times 10^{-5}$). Although this gene set was also enriched in blood ($P=1.5 \times 10^{-6}$) and all tissues combined ($P=1.4 \times 10^{-4}$), it was associated with positive age-related CpGs instead indicating that some age-related gene sets can be impacted by negative and positive age-related CpGs, potentially influencing different members of the set or perhaps having opposing transcriptional outcomes resulting from methylation. Another highly-relevant example of this is the observation concerning mitochondrial function. While hypomethylated age-related cytosines in brain, cortex, and muscle are enriched for numerous mitochondria-related genes; in blood and skin, however, these are enriched for positive age-related cytosines (**Extended Data Table 3**).

Overlap of age-related cytosines with human traits and diseases

To uncover potential correlation between age-related cytosines and known human traits, the proximal genomic regions of the same top 1,000 positively and 1,000 negatively associated CpGs were overlaid with the top 5% of genes that were associated with numerous human traits identified by GWAS. At threshold of $P<5.0\times10^{-4}$, overlaps were found with genes associated with longevity, Alzheimer's, Parkinson's and Huntington's disease, dementia, epigenetic age acceleration, age at menarche, leukocyte telomere length, inflammation, mother's longevity, metabolic diseases, obesity (fat distribution, body-mass index), etc. (**Extended Data Fig. 5**, **Supplementary Tables 6.1-6.7)**; many of which are associated with advancing age.

Development of universal pan-tissue epigenetic clocks of age across mammals

Having identified age-related cytosines shared across mammalian species and tissues, we proceeded to use them to develop universal mammalian epigenetic age clocks. We developed three universal mammalian age-estimators, which differ with respect to output. The first, *universal naïve clock (Clock 1)* directly correlates DNA methylation profile to chronological age. To allow biologically meaningful comparisons between species with very different life-spans, we developed a second universal clock that defines individual age relative to the maximum lifespan of its species; generating *relative age* estimates between 0 and 1. As the accuracy of this *universal relative age clock (Clock 2)* could be compromised in species for which knowledge of maximum lifespan is unavailable, a third universal clock was developed, which omits maximum lifespan and uses instead average age at sexual maturity. Age at sexual maturity was chosen as species characteristics since it correlates strongly with maximum lifespan on the log scale (Pearson correlation $r=0.82$, $p=6\times10^{-183}$ across all mammalian species in AnAge**)**. This third clock is referred to as the *universal log-linear transformed age clock (Clock 3)*.

Performance of universal epigenetic clocks across species

We employed two different strategies for evaluating the accuracy of the clocks. First, leave-one-fraction-out (LOFO) cross-validation analysis randomly divided the data set into 10 fractions, each of which contained the same proportions of species and tissue types, and a different fraction is left out for validation at each iteration of analysis. Second, leave-one-species-out analysis (LOSO) was similarly cross-validated with the omission of a species at each iteration.

According to LOFO cross-validation, the epigenetic clocks were remarkably accurate (r>0.96), with a median error of less than 1 year and a median relative error of less 3.5 percent **(Figs. 2a, 3a-b, Extended Data Table 4).** According to the LOSO evaluation, the clocks reached age correlations up to r=0.94 (**Extended Data Table 4**). The median correlation (and MAE) across species was as strong with either LOFO or LOSO evaluations. For some species such as bowhead whales, however, epigenetic age as predicted by the naïve clock accords poorly with chronological age (**Fig. 2b**). We investigated and ascertained that the mean difference between LOSO DNAmAge and chronological age is negatively correlated with maximum lifespan (r=-0.57, p=3x10^{-6}) and age at sexual maturity (r= -0.5, p=6.4x10^{-5}) of the species (**Fig. 2c-d**). Here, the strength of clock 2 comes to fore as it is not affected by maximum lifespan, which was incorporated into it during its construction. Clock 2 and clock 3 achieve a correlation of r=0.96 and r=0.95 between DNAm and observed relative age, respectively (**Fig. 3d,e**). Both of these clocks present comparably accurate LOFO estimates in numerous tissue types in 58 species (**Supplementary Table 8.2)**, with a representation in **Fig. 3g-i** of LOFO Clock 2 correlations for humans (r=0.961, 19 tissues), mice (r=0.954, 25 tissues), and bottlenose dolphins (r=0.95, 2 tissues). While the clock accurately predicted the age for one mysticete species, the humpback whale and all other mammalian species, the ages of bowhead samples were sometimes underestimated (species index 3.11 in **Fig. 3a,b**). This may simply reflect the inaccuracy of the age estimations used for bowhead whales, which were aged using the aspartic acid racemization rate. These clocks are similarly accurate with LOSO age-estimates between evolutionarily distant species

(Supplementary Table 9.2) including dogs (r=0.917, MAE=1.3), Savanna elephants (r=0.962, MAE<3 years), and flying foxes (r=0.982, MAE=1.2) (**Fig. 3j-l**). Such high predictive accuracy of LOSO analysis demonstrates that these universal clocks are applicable to mammalian species that are not part of the training data (**Supplementary Tables 9.1, 9.2)**. The three universal clocks performed just as well in 63 species, for which there were fewer than 15 samples (r~0.9, MAE~1 year, **Extended Data Fig. 6a-c**), showing very strong correlation between estimated and actual relative age (r=0.92, **Extended Data Fig. 6d).**

With regards to marsupials, we encountered two limitations. First, less than half of the eutherian CpGs apply to marsupials [2]. Second, there were only seven marsupial species in our data set with total sample size N=162. These limitations notwithstanding, we were still able to construct a fourth universal clock for estimating relative age in marsupials (age correlation, r=0.88, med.Cor=0.87 in **Fig. 3c,f**).

Performance of universal epigenetic clocks across tissues

As the epigenome landscape varies markedly across tissue types [9,10], we assessed tissue-specific accuracy of clock 2 for relative age (r=0.96, **Fig. 3d**). Of the 33 distinct tissue types, the median correlation is 0.94 and median MAE for relative age is 0.026 (**Supplementary Table 8.3)**. There was high age-correlation with whole brain (r=0.987), cortex (r=0.972), hippocampus (r=0.964), striatum (r=0.956), cerebellum (r=0.975), spleen (r=0.981), and kidney (r=0.979) (**Fig. 4)**. Blood and skin also exhibited similarly high estimates of relative age correlations across different species: blood (r=0.958, MAE=0.018, 74 species) and skin (r=0.948, MAE=0.026, 56 species) (**Fig. 4i,n)**.

DISCUSSION

The universality of aging across all mammalian species has engendered speculations of its cause, with the predominant notion that random damage to cellular constituents underlies this process. The ability

to accurately estimate ages of mammals by virtue of their methylation profiles however, introduces the likelihood of a largely deterministic process. We investigated this question by generating an unprecedentedly large set of DNA methylation profiles from over 121 eutherian species and 7 marsupial species, from which an unambiguous feature emerged. Genes that are proximal to age-related CpGs, overwhelmingly represent those involved in the process of development, such as *HOX* and *PAX*. This is consistent with enrichment of these cytosines in target sites of PRC2 and bivalent chromatin domains, which control expression of *HOX* and other developmental genes in all vertebrates and beyond. It appears therefore, that aging is hard-wired into life through processes associated with development.

A large body of literature connects growth/development to aging starting with the seminal work by Williams 1957 [11]. This connection is also apparent when Yamanaka factor-mediated reversion of adult cells to embryonic stem cells is accompanied by resetting of their age to prenatal epigenetic age, matching their development stage [1]. Therefore, methylation regulation of the genes involved in development (during and after the developmental period) may constitute a key mechanism linking growth and aging. The universal epigenetic clocks demonstrate that aging and development are coupled and share important mechanistic processes that operate over the entire lifespan of an organism.

Other notable age-related genes and processes that were uncovered include *LHFPL4* and *LHFPL3* whose reported function in synaptic clustering of GABA receptors does not immediately present an obvious connection to aging across all tissues. However, the extremely strong correlation of CpGs near these paralogues (located on separate chromosomes) with age argues strongly for their role in the aging process. The *LARP1* gene ranks first in liver and second across all tissues for hypomethylation with age and encodes a protein that regulates translation of downstream targets of mTOR, which has very well-documented links with aging and longevity. The implication of circadian rhythm genes, exclusively in aging brain tissues, reveals tissue-specific changes that occur in parallel with universal

developmental ones. Furthermore, involvement of circadian rhythm genes in aging echoes recent observations in mice [4].

The implication of multiple genes related to mitochondrial function supports the long-argued importance of this organelle in the aging process. It is also important to note that many of the identified genes are implicated in a host of age-related pathologies and conditions, bolstering the likelihood of their active participation in, as opposed to passive association with, the aging process.

Future elucidation of how development is mechanistically connected to aging will be aided by the universal mammalian clocks. The leave-one-species-out cross validation analysis demonstrates that these clocks generalize very well to mammalian species that were not part of the training set. The ability to construct universal mammalian epigenetic clocks that can accurately predict the age of animals and tissues that were not part of the training set fulfils Popper's dictum of falsifiability, which requires that a theory make testable predictions on the basis of which it can be refuted. The epigenetic clocks presented here, built on the universality of mammalian aging, pass this test with remarkable ease and accuracy.

METHODS

Tissue samples

The tissue samples are described in the Supplement and related citations as listed in **Supplementary Information, Note 1)**.

Quality controls for establishing universal clocks

We generated two variables to guide the quality control (QC) of the study samples; the first being a variable indicating the confidence (0 to 100%) in the chronological age estimate of the sample. For example, a low confidence was assigned to samples from wild animals whose ages were estimated based on body length measurements. The epigenetic clocks were trained and evaluated in tissue

samples whose confidence exceeded 90% (>=90%). The second quality control variable was an indicator variable (yes/no) that flagged technical outliers or malignant (cancer) tissue. Since we were interested in "normal" aging patterns we excluded tissues from preclinical studies surrounding anti-aging or pro-aging interventions.

Species characteristics

Species characteristics such as maximum lifespan (maximum observed age), age at sexual maturity, and gestational length were obtained from an updated version of the Animal Aging and Longevity Database [12] (AnAge, http://genomics.senescence.info/help.html#anage). To facilitate reproducibility, we have posted this modified/updated version of AnAge in Supplementary Data.

Meta analysis EWAS of age

Each CpG was regressed on chronological age in each stratum formed by species/tissue. We limited the analysis to strata that contained at least 15 observation. This correlation test resulted in a Student t-test statistic (denoted as Z statistic). We computed two different meta analysis statistics. The first approach (Stouffer's method) combined the P values (and corresponding Z statistics) of different species/tissue strata using the Metal software [13] (Methods). Stouffer's meta analysis is attractive since it allowed us to calculate meta analysis p values for each CpG. The second meta analysis approach simply calculated the median Z statistic across the strata. We found that this approach is attractive for pre-filtering CpGs in the training sets of universal clocks. We emphasize that this pre-filtering approach did not include any of the test data. In each training set we pre-filtered top 4000 CpGs before modeling the clocks. For clocks 1 and 2, we used the median Z statistics; for clock 3, we used the "rankPvalue" R function from the WGCNA R package applied to age correlations; for clock 4, we used roughly 14,500 CpGs that mapped to the genomes of opossum and Tasmanian devil.

Meta analysis for EWAS of age

We carried out two methods to combine EWAS results across species and tissues, as described below.

Two-stage meta analysis in conjunction with Stouffer's method

Our meta analysis of age combined correlation test statistics calculated in 133 different species-tissue strata (from 58 species) with a minimal sample size of 15 (N≥15, Supplementary Table 1.4). In the first stage, we combined the EWAS results across tissues within the same species to form species specific meta-EWAS results. In the second stage, we combined the total of 58 species EWAS results to form a final meta-EWAS of age. All the meta analyses in both stages were performed by the unweighted Stouffer's method, as conducted in METAL[13].

Stratification of age groups

To assess whether the age related CpGs in young animals relate to those in old animals, we split the data into 3 age groups: young age (age < 1.5* age at sexual maturity, ASM), middle age (age between 1.5 and 3.5 ASM), and old age group (age ≥ 3.5 ASM). The threshold of sample size in species-tissue was relaxed to N≥10. The age correlations in each age group were meta analyzed using the above mentioned two-stage meta analysis approach.

Brain EWAS

Analogously, we applied the two approaches to brain EWAS results; more than 900 brain tissues from human, vervet monkey, mice, olive bamboo, brown rat, and pig species across cerebellum, cortex, hippocampus, hypothalamus, striatum, subventricular zone (SVZ), and whole brain.

EWAS of single tissue

One-stage unweighted Stouffer's method and Median Z score were also applied to EWAS results from cerebellum and cortex, respectively. Similarly, we carried out meta-analysis EWAS of blood, liver, muscle, and skin. Blood EWAS results were combined across 7 families including 367 tissues from humans, 565 from dogs, 170 from mice, 36 from killer whales, 137 from bottlenose dolphins, 83 from

Asian elephants, etc. Skin EWAS results were combined across 5 families including 95 from bowhead whales, 638 tissues from 19 bat species, 180 from killer whales, 105 from naked mole rats, 72 from humans, etc. Liver EWAS results were combined across four families including 583 mice, 97 from humans, 48 from horses, etc. Muscle EWAS results were combined across four families including 24 from evening bats, 57 from humans, and 19 from naked mole rats, etc. Cerebellum EWAS results were combined across Primates and Rodentia including 46 from humans. Another 46 cerebral cortex tissues profiled in the same human individuals were included in the cortex EWAS, in which the meta analysis was also combined across Primates, Rodentia, and a third Order: 16 pigs from Artiodactyla. Details surrounding the different species-tissue strata are presented in **Supplementary Table 2**.

We used the R gmirror function to depict mirror image Manhattan plots.

GREAT analysis

We applied the GREAT analysis software tool[6] to the top 1000 hypermethylated and the top 1000 hypomethylated CpGs from EWAS of age. GREAT implemented foreground/background hypergeometric tests over genomic regions where we input all of the 37k CpG regions of our mammalian array as background and the genomic regions of the 1000 CpGs as foreground. This yielded hypergeometric p-values not confounded by the number of CpGs within a gene. We performed the enrichment based on the settings (assembly: Hg19, Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 50 kb) for about 76,290 gene sets associated with GO terms, MSigDB (including gene sets for upstream regulators), PANTHER, KEGG pathway, disease ontology, gene ontology, human and mouse phenotypes. We report the gene sets with FDR <0.05 and list nominal hypergeometric P-values, FDR and Bonferroni corrected P-values.

EWAS-GWAS based overlap analysis

Our EWAS-GWAS based overlap analysis related the gene sets found by our EWAS of age with the gene sets found by published large-scale GWAS of various phenotypes, across body fat distribution,

lipid panel outcomes, metabolic outcomes, neurological diseases, six DNAm based biomarkers, and other age-related traits (**Supplementary Information**, **Note 3**). A total of 69 GWAS results were studied. The six DNAm biomarkers included four epigenetic age acceleration measures derived from 1) Horvath's pan-tissue epigenetic age adjusted for age-related blood cell counts referred to as intrinsic epigenetic age acceleration (IEAA) [1,14], 2) Hannum's blood-based DNAm age[15]; 3) DNAmPhenoAge [16]; and 4) the mortality risk estimator DNAmGrimAge [17], along with DNAm based estimates of blood cell counts and plasminogen activator inhibitor 1(PAI1) levels[17]. For each GWAS result, we used the MAGENTA software to calculate an overall GWAS *P*-value per gene, which is based on the most significant SNP association *P*-value within the gene boundary (+/- 50 kb) adjusted for gene size, number of SNPs per kb, and other potential confounders [18]. We pruned in the genomic regions of GWAS genes present in the mammalian array. For each EWAS results, we studied the genomic regions from the top 1000 CpGs hypermethylated and hypomethylated with age, respectively. To assess the overlap with a test trait, we selected the top 5 % genes for each GWAS trait and calculated one-sided hypergeometric P values based on genomic regions (as detailed in [19,20]). The number of background genomic regions in the hypergeometric test was based on the overlap between the entire genes in a GWAS and the entire genomic regions in our mammalian array. We highlighted the GWAS trait when its hypergeometric P value reached $5x10^{-4}$ with EWAS of age in any tissue type.

Association of *LHFPL* gene expression with chronological age in human and mouse

To study if *LHFPL4* or *LHFPL3* play a role in age-related transcriptional changes surrounding nearby genes, we analyzed several transcriptomic data across multiple tissues and species. In humans, our analysis leveraged gene expression studies from 1) GTEx project, 2) two gene expression data studied in [19] (GEO datasets from studies [21,22]) and 3) the summary data across three studies in Isildak et al.[3].

for investigating age-related brain expression in developmental (age <20) and aging (age ≥20) periods. In mice, we analyzed the summary data from The Tabula Muris Consortium [4] which generated single cell RNA seq data from 23 mouse tissues across the lifespan.

Three universal mammalian clocks for eutherians

We applied elastic net regression models to establish three universal mammalian clocks for estimating chronological age across all tissues in eutherians. The three elastic net regression models corresponded to different outcome measures described in the following: 1) log transformed chronological age: log(Age+2) where an offset of 2 years was added to avoid negative numbers in case of prenatal samples, 2) –log(-log(RelativeAge)) and 3) log-linear transformed age. DNAm age estimates of each clock were computed via the respective inverse transformation. Age transformations used for building universal clocks 2-4 incorporated three species characteristics: gestational time (GT), age at sexual maturity (ASM), and maximum lifespan ($maxAge$). All of these species variables surrounding time are measured in units of years. The details for each species are presented in Supplementary Data.

<u>Loglog transformation of Relative Age for clock 2</u>

Our measure of relative age leverages gestation time (*GestationT*) and maximum lifespan. We define relative age ($RelativeAge$) and apply Log-log transformation as the following:

$$RelativeAge = \frac{Age + GestationT}{MaxAge + GestationT} \ (1)$$

$$LoglogAge = -\log(-\log(RelativeAge)) \ (2)$$

By definition, $RelativeAge$ is between 0 to 1 and $LoglogAge$ is positively correlated with age. Universal clock 2 predicts $LoglogAge$ and next applies an inverse transformation to estimate DNAmAge:

$$DNAmAge = \exp(-\exp(-LoglogAge)) * (MaxAge + GestationT) - GestationT \ (3)$$

All species characteristics (e.g. MaxAge, gestational time) come from our updated version of AnAge. We were concerned that the uneven evidence surrounding the maximum age of different species could bias our analysis. While billions of people have been evaluated for estimating the maximum age of humans (122.5 years), the same cannot be said for any other species. To address this concern, we made the following admittedly ad-hoc assumption: the true maximum age of non-human species is 30% higher than that reported in AnAge. Therefore, we multiplied the reported maximum lifespan of non-human species by 1.3. Our predictive models turn out to be highly robust with respect to this assumption (data not shown).

Transformation based on log-linear age for clock 3

Our measure of log-linear age leverages age at sexual maturity (ASM). The transformation has the following properties: takes the logarithmic form when age is less than ASM; takes the linear form when age is greater than ASM; continuously differentiable at ASM.

First, we define a ratio of the age relative to ASM as the following:

$$RelativeAdultAge = \frac{Age+1.5}{ASM+1.5} \quad (4),$$

where the offset of 1.5 years ensures the $RelativeAdultAge$ is positive. To accommodate a faster growth rate in young age, we apply a log-linear transformation on $RelativeAdultAge$ using the function $f(x)$ that was originally proposed for the human pan tissue clock [1]:

$$y = f(x) = \begin{cases} x - 1\,, x \geq 1 \\ \log(x)\,, x < 1 \end{cases} \quad (5)$$

$$f^{-1}(y) = \begin{cases} y + 1\,, y \geq 0 \\ \exp(y)\,, y < 0 \end{cases} \quad (6)$$

This transformation ensures continuity and smoothness at the change point $x = 1$. In our study, the argument x is the ratio $RelativeAdultAge$. Hence, we denote a sample at young age if its ratio $RelativeAdultAge$ is less than 1. Our log-linear age ($LoglinearAge$) in clock 3 is expressed below:

$$LoglinearAge = \begin{cases} RelativeAdultAge - 1, RelativeAdultAge \geq 1 \\ \log(RelativeAdultAge)\,, RelativeAdultAge < 1 \end{cases} \quad (7).$$

Analogously, clock 3 predicts $LoglinarAge$ and applies inverse transformation to estimate DNAmAge as below.

$$DNAmAge = \begin{cases} ASM + LoglinearAge * (ASM + 1.5)\,, LoglinearAge \geq 0 \\ \exp(LoglinearAge) * (ASM + 1.5) - 1.5, LoglinearAge < 0 \end{cases} \quad (8)$$

Marsupial clock

For marsupial clock, we used 162 samples across 7 species. We applied elastic net regression to the outcome measure $LoglogAge = -\log(-log(RelativeAge))$ as described in formula (1) & (2). For accessing the accuracy of the clock, we only used LOFO cross validation (with 5 fractions) since the majority of the samples came from opossums (N=100) and Tasmanian devils (N=41**, Supplementary Table 1.1)**.

We used a different pipeline for normalizing the methylation data for marsupials because many CpGs in other mammals did not map to the marsupial genome. The marsupial clock was trained on the basis of roughly 14500 cytosines that mapped to both Tasmanian devil and opossums.

Elastic net regression

We applied the elastic net regression models to train all samples that selected 1000 to 2000 CpGs for clocks 1-3 and 30 CpGs for the marsupial clock. To assess the accuracy of the elastic net regression models, we used leave-one-fraction-out (LOFO) and leave one-species-out (LOSO) cross validation. In LOFO, we randomly split the entire dataset into 10 fractions each of which had the same distribution in species and tissue types. Each penalized regression model was trained in 9 fractions but evaluated in the 10th left out fraction. After circling through the 10 fractions, we arrived at LOFO predictions which were subsequently related to the actual values.

The LOSO cross validation approach trained each model on all but one species. The left out species was used a test set. The LOSO approach was used to assess how well the penalized regression models generalize to species that were not part of the training data. To ensure unbiased estimates of accuracy,

all aspects of the model fitting (including pre-filtering of the CpG) were only conducted in the training data in both LOFO and LOSO analysis. Elastic net regression in the training data was implemented by setting the glmnet model parameter alpha to 0.5. Ten-fold cross validation in the *training* data was used to estimate the tuning parameter lambda. For computational reasons, we fitted the glmnet model to the top 4000 CpGs with the most significant median Z score (age correlation test) in the training data. To accommodate different samples sizes of the species we used weighted regression as needed where the weight was the inverse of square root of species frequency or 1/20 (whichever was higher). The final versions of the different universal clocks used all available data.

Statistics for performance of model prediction

To validate our model, we used DNAm age estimates from LOFO and LOSO analysis, respectively. At each type of estimates, we performed Pearson correlation coefficients and computed median absolute difference (MAE) between DNAm based and observed variables across all samples. Correlation and MAE were also computed at species level, limited to the subgroup with samples N>=15 (within a species or within a species-tissue category). We reported the medians for the correlation estimates (**med.Cor**) and the medians for the MAE estimates (**med.MAE**) across species, respectively. Analogously, we repeated the same analysis at species-tissue level, limited to the subgroup with sample N >=15 (within a specie-tissue category).

URLs

AnAge, http://genomics.senescence.info/help.html#anage

UCSC genome browser: http://genome.ucsc.edu/index.html

ACKNOWLEDGEMENTS and **FUNDING**

This work was mainly supported by the Paul G. Allen Frontiers Group (SH).

Data availability

The data will be posted on Gene Expression Omnibus.

CONTRIBUTIONS

Ake T. Lu, Zhe Fei, Caesar Li, Joseph Zoller, SH developed the universal clocks. ATL, Amin Haghani, Charles Breeze, Michael Thompson, Matteo Pellegrini, Wanding Zhou, SH carried out additional bioinformatics analyses. Adriana Arneson, Jason Ernst, SH designed the mammalian methylation array. ATL, Ken Raj, SH drafted the first version of the article. The remaining authors contributed precious tissues or DNA samples or helped with the data generation process. All authors helped with editing the article and data interpretation. SH conceived of the study.

COMPETING INTERESTS

SH is a founder of the non-profit Epigenetic Clock Development Foundation which plans to license several patents from his employer UC Regents. These patents list SH, JE, AA as inventors. The other authors declare no conflicts of interest.

CORRESPONDING AUTHOR

Correspondence to Steve Horvath (shorvath@mednet.ucla.edu)

Figure Legends.

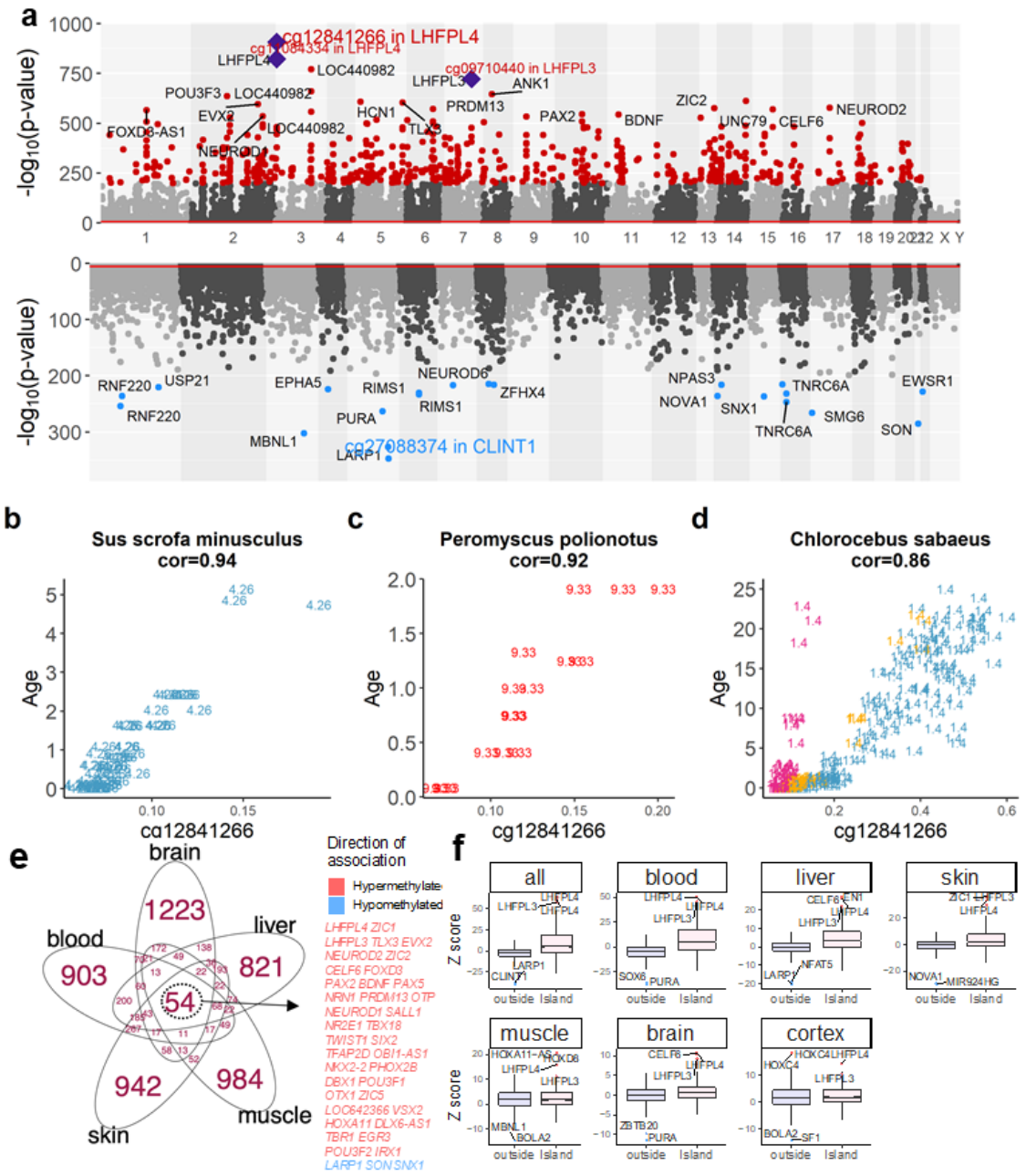

Figure 1. Meta-analysis of chronological age across species and tissues. a, Meta-analysis p-value (-log base 10 transformed) versus chromosomal location (x-axis) according to human genome assembly 38 (Hg38). The upper and lower panels of the Manhattan plot depict CpGs that gain/lose methylation with age. CpGs colored in red and blue exhibit highly significant ($P<10^{-200}$) positive and negative age correlations, respectively. The most significant CpG (cg12841266, $P=9.3x10^{-913}$) is located in exon 2 of the *LHFPL4* gene in humans and most other mammalian species, followed by cg11084334 ($P=1.3x10^{-827}$). These two CpGs and cg097720($P=4.3x10^{-725}$) located in the paralog gene LHFPL3 are marked in purple diamond points. Scatter plots of cg12841266 (in *x*-axis) versus chronological age in **b**, mini pigs (*Sus scrofa minusculus*), **c**, Oldfield mouse (*Peromyscus polionotus)*, and **d**, vervet monkey (*Chlorocebus aethiops sabaeus*), respectively. Tissue samples are labelled by the mammalian species index and colored by tissue type as detailed in Supplemental Table 1s. Panels e-g: annotations of the top 1000 hypermethylated and hypomethylated CpGs listed in the EWAS meta-analysis across all (results in panel a), brain, blood, liver, and skin tissues, respectively. **e**, the Venn diagram displays the overlap of age-associated CpGs across different organs, based on EWAS of the top 1000 hypermethylated/hypomethylated CpGs. We list all 36 genes that are proximal to the 54 age-associated CpGs common across all organs in the Venn diagram. **f,** the bar plots depicts the associations of the EWAS results (meta Z scores) with CpG islands (inside/outside) in different tissue types. We list top genes for each bar. **g**, Selected results from GREAT enrichment analysis. The color gradient is based on -log10 (hypergeometric P value). The size of the points reflects the number of common genes.

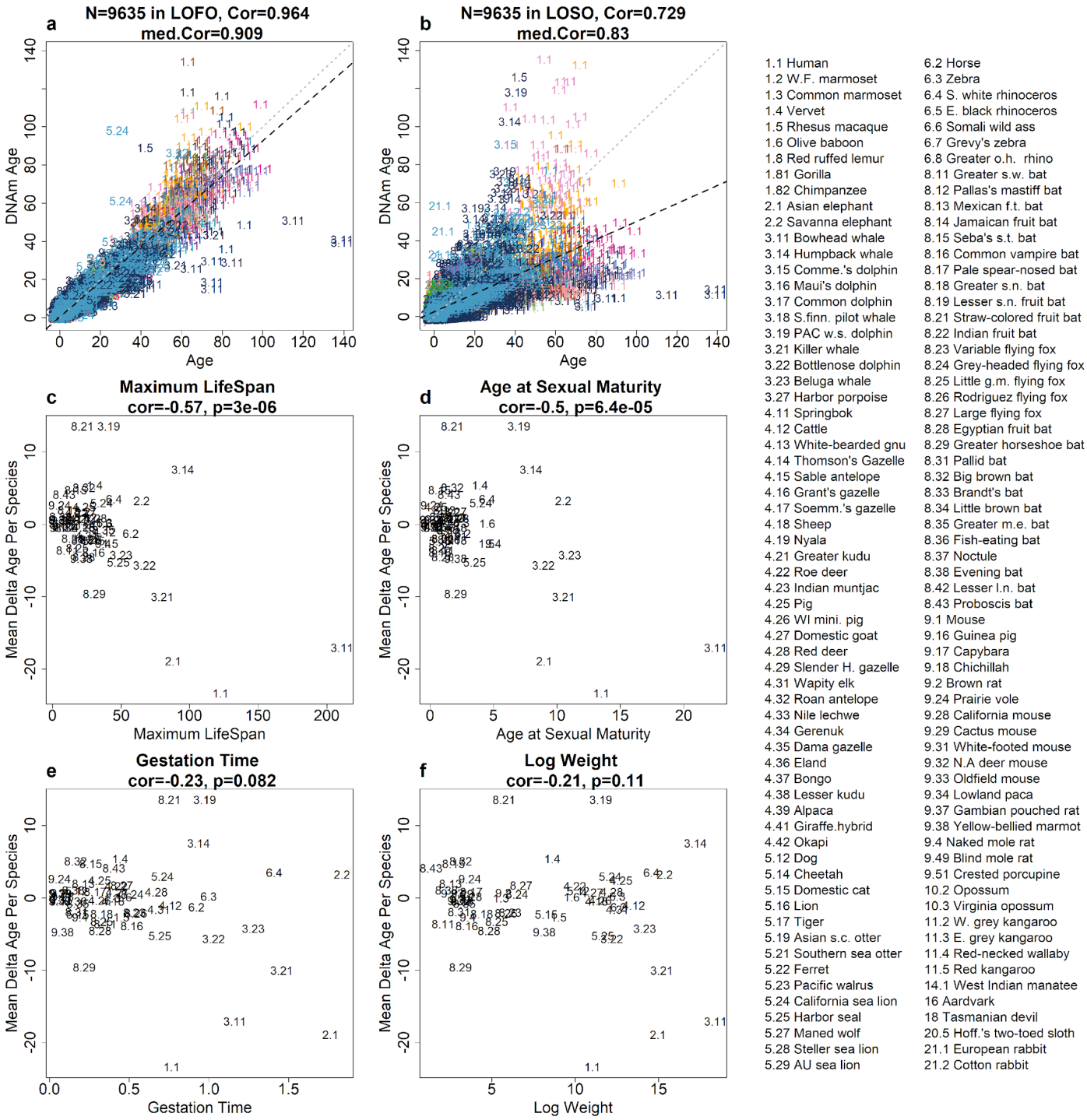

Figure 2. Naïve universal clock for log-transformed age. a, b, Chronological age (x-axis) versus DNAmAge estimated using **a,** leave-one-fraction-out (LOFO) **b,** leave-one-species-out (LOSO) analysis**.** The grey and black dashed lines correspond to the diagonal line (y=x) and the regression line, respectively. Each dot (tissue sample) is labelled by the mammalian species index (legend). The species index corresponds to the phylogenetic order, e.g. 1=primates, 2=elephants (Proboscidea), 3=cetaceans etc. The number after the decimal point denotes the individual species within the phylogenetic order. Points are colored according to designated tissue color (**Supplemental Table 1.3**). The heading of each panel reports the Pearson correlation (cor) across all samples. The med.Cor (or med.MAE) is the median across species that contain 15 or more samples. **c-f**, Delta age denotes the difference between the LOSO estimate of DNAm age and chronological age. The scatter plots depict

mean delta age per species (y-axis) versus **c**, maximum lifespan observed in the species, **d**, average age at sexual maturity **e**, gestational time (in units of years), and **f**, (log-transformed) average adult weight in units of grams.

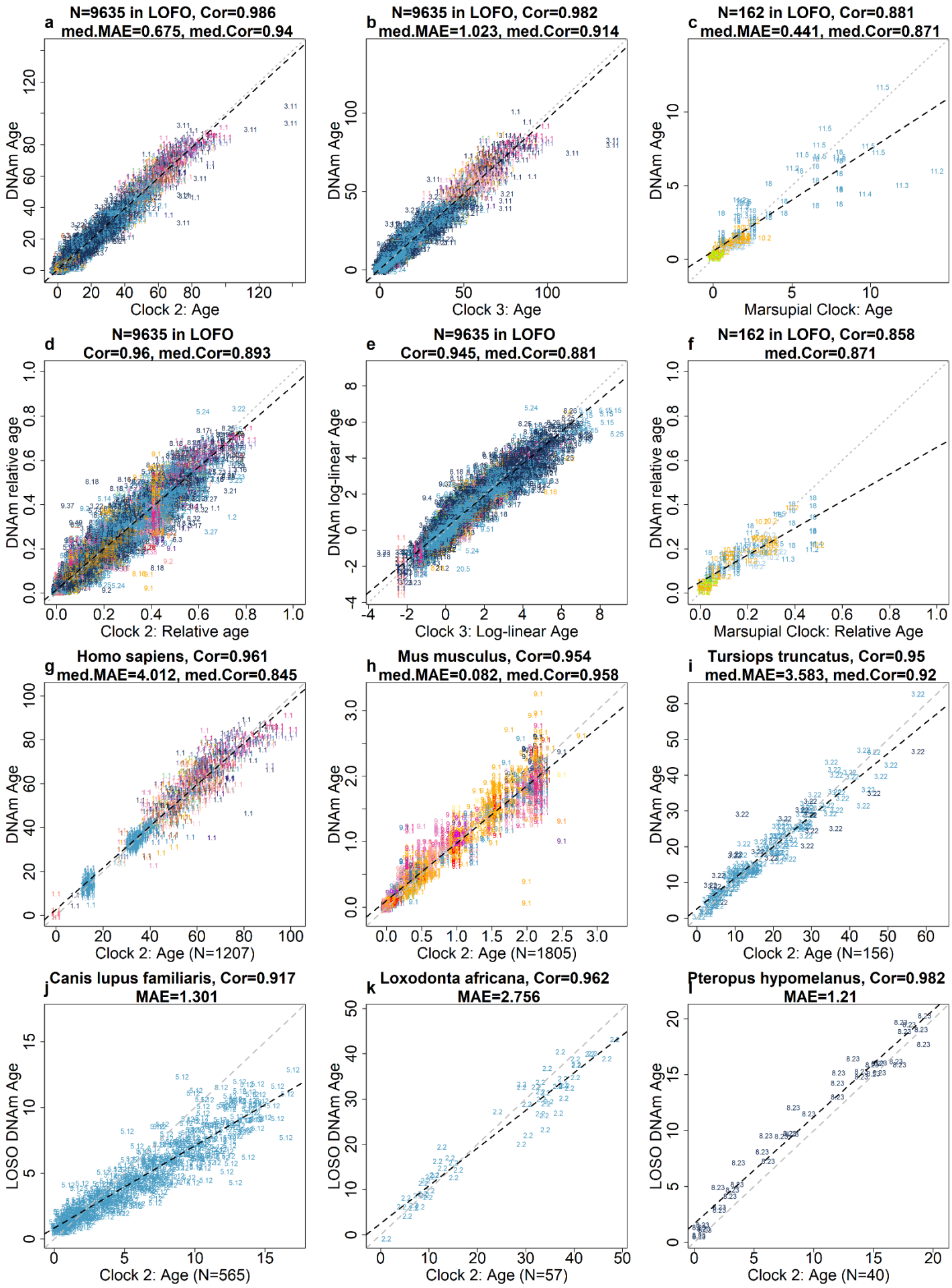
a
N=9635 in LOFO, Cor=0.986
med.MAE=0.675, med.Cor=0.94
DNAm Age
Clock 2: Age
b
N=9635 in LOFO, Cor=0.982
med.MAE=1.023, med.Cor=0.914
DNAm Age
Clock 3: Age
c
N=162 in LOFO, Cor=0.881
med.MAE=0.441, med.Cor=0.871
DNAm Age
Marsupial Clock: Age
d
N=9635 in LOFO
Cor=0.96, med.Cor=0.893
DNAm relative age
Clock 2: Relative age
e
N=9635 in LOFO
Cor=0.945, med.Cor=0.881
DNAm log-linear Age
Clock 3: Log-linear Age
f
N=162 in LOFO, Cor=0.858
med.Cor=0.871
DNAm relative age
Marsupial Clock: Relative Age
g
Homo sapiens, Cor=0.961
med.MAE=4.012, med.Cor=0.845
DNAm Age
Clock 2: Age (N=1207)
h
Mus musculus, Cor=0.954
med.MAE=0.082, med.Cor=0.958
DNAm Age
Clock 2: Age (N=1805)
i
Tursiops truncatus, Cor=0.95
med.MAE=3.583, med.Cor=0.92
DNAm Age
Clock 2: Age (N=156)
j
Canis lupus familiaris, Cor=0.917
MAE=1.301
LOSO DNAm Age
Clock 2: Age (N=565)
k
Loxodonta africana, Cor=0.962
MAE=2.756
LOSO DNAm Age
Clock 2: Age (N=57)
l
Pteropus hypomelanus, Cor=0.982
MAE=1.21
LOSO DNAm Age
Clock 2: Age (N=40)

Figure 3. Universal clocks for transformed age across mammals. The figure displays universal clock 2 (Clock 2) estimates of relative age, universal clock 3 (Clock 3) estimates of log-linear transformation of age and marsupial clock (Marsupial Clock) estimates of relative age of eutherian and marsupial samples respectively. Relative age estimation incorporates maximum lifespan and gestational age, and assumes values between 0 and 1. Log-linear age is formulated with age at sexual maturity and gestational time. The DNAm estimates of age (y axes) of (**a**) and (**b**) are transformation of relative age (Clock 2 and Marsupial Clock) or log-linear age (Clock 3), into units of years. **a-f**, Age estimated via leave-one-fraction-out (LOFO) cross-validation for Clock 2 (**a,d**), Clock 3 (**b,e**) and Marsupial Clock (c,f). **g-i**, Age estimated via LOFO cross-validation in Clock 2. **j-l**, age estimated via leave-one-species-out (LOSO) cross-validation for Clock 2. We report Pearson correlation coefficient estimates. Median correlation (med.Cor) and median of median absolute error (med.MAE) are calculated across species (**a-f**) or across species-tissue (**g-i**). Each sample is labelled by mammalian species index and marked by tissue color (**Fig. 2**, **Supplementary Table**s 1.1-1.2).

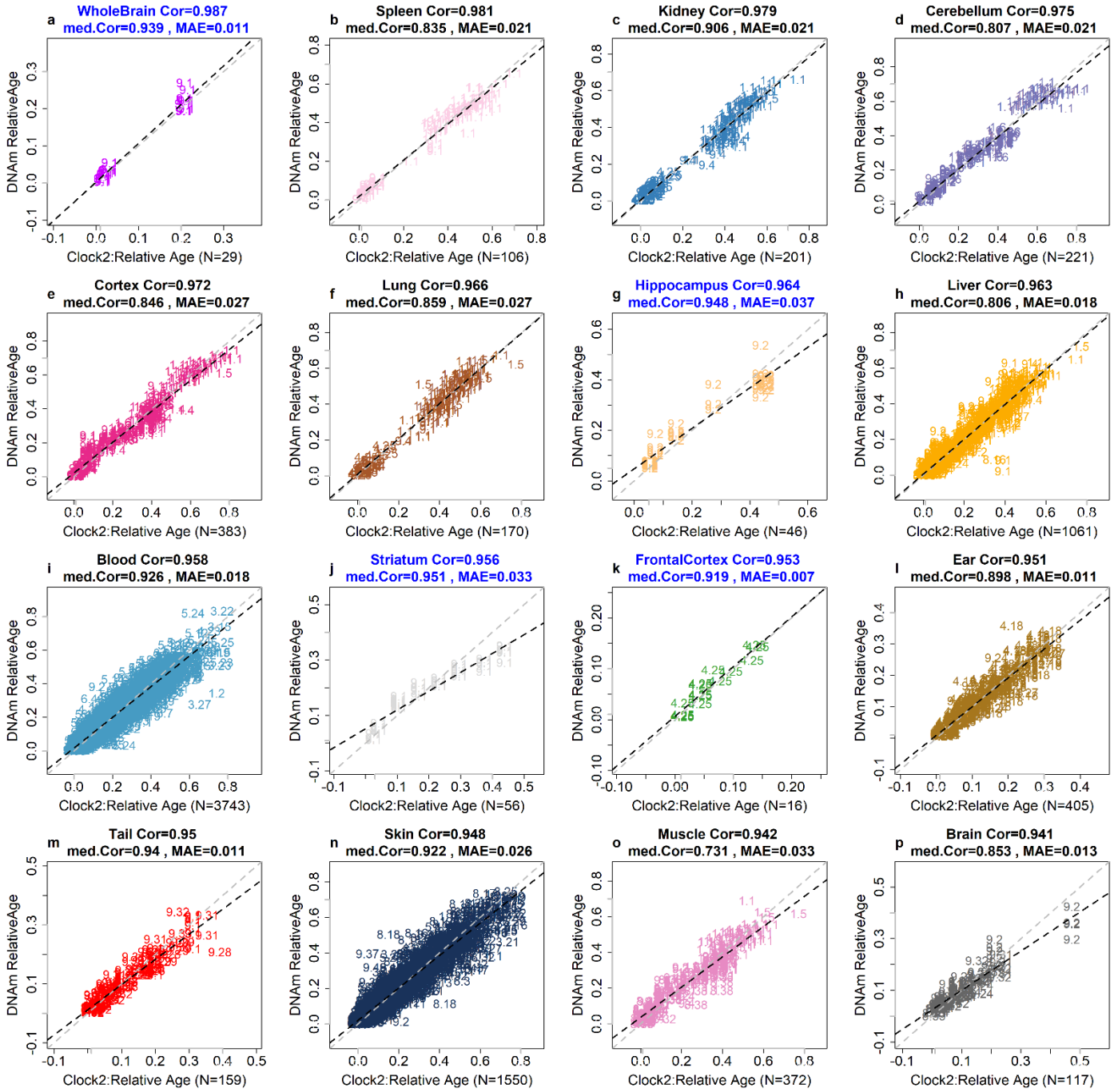

Figure 4. Universal clock for relative age applied to specific tissues. The specific tissue or cell type is reported in the title of each panel. DNA methylation based estimates of relative age (y-axis) versus actual relative age (x-axis). Each dot presents a tissue sample colored by tissue and labelled by mammalian species index (Supplementary Tables 1.2-1.3). The analysis is restricted to tissues with at least 15 samples available. Leave-one-folder-out cross-validation (LOFO) was used to arrive at unbiased estimates of predictive accuracy measures: median absolute error (MAE) and age correlation based on relative age. "Cor" denotes the Pearson correlation coefficient based on all available samples. "med.Cor" denotes the median values across all species for which at least 15 samples were available. Title is marked in blue if a tissue type was collected from a single species.

REFERENCE

1 Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115, doi:10.1186/gb-2013-14-10-r115 (2013).

2 Arneson, A. *et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *bioRxiv*, 2021.2001.2007.425637, doi:10.1101/2021.01.07.425637 (2021).

3 Işıldak, U., Somel, M., Thornton, J. M. & Dönertaş, H. M. Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific Reports* **10**, 4080, doi:10.1038/s41598-020-60998-0 (2020).

4 A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1 (2020).

5 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, doi:10.1038/nbt.1630 (2010).

6 Rakyan, V. K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research* **20**, 434-439, doi:10.1101/gr.103101.109 (2010).

7 Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research* **20**, 440-446, doi:10.1101/gr.103606.109 (2010).

8 Petkovich, D. A. *et al.* Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions. *Cell Metab* **25**, 954-960 e956, doi:10.1016/j.cmet.2017.03.016 (2017).

9 Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotech* **28**, 1045-1048 (2010).

10 Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* **6**, e22, doi:10.1371/journal.pbio.0060022 (2008).

11 Williams, G. C. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* **11**, 398-411, doi:10.2307/2406060 (1957).

12 de Magalhaes, J. P., Costa, J. & Church, G. M. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J Gerontol A Biol Sci Med Sci* **62**, 149-160 (2007).

13 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

14 Horvath, S. *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol* **17**, 171, doi:10.1186/s13059-016-1030-0 (2016).

15 Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**, 359-367, doi:10.1016/j.molcel.2012.10.016 (2013).

16 Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, doi:10.18632/aging.101414 (2018).

17 Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303-327, doi:10.18632/aging.101684 (2019).

18 Segre, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**, doi:10.1371/journal.pgen.1001058 (2010).

19 Lu, A. T. *et al.* Genetic architecture of epigenetic and neuronal ageing rates in human brain regions. *Nat Commun* **8**, 15353, doi:10.1038/ncomms15353 (2017).

20 Lu, A. T. *et al.* Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum. *Nat Commun* **7**, 10561, doi:10.1038/ncomms10561 (2016).

21 Gibbs, J. R. *et al.* Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *PLoS Genet* **6**, e1000952 (2010).

22 Hernandez, D., Nalls, M., Gibbs, J. & et al. Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum Mol Genet* **20**, 1164–1172 (2011).