Three universal pan-mammalian clocks

We applied elastic net regression models to establish three universal mammalian clocks for estimating chronological age across all tissues (n=11,754 from 185 species) in eutherians (N=11439 from 176 species), marsupials (n=210 from 9 species) and monotremes (n=15 from 2 species). 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 and 3 incorporated three species characteristics: gestational time ($GestationT$), age at sexual maturity (ASM), and maximum lifespan ($MaxLifespan$). All of these species variables surrounding time are measured in units of years.

Loglog transformation of Relative Age for clock 2

Our measure of relative age leverages gestation time and maximum lifespan. We define relative age ($RelativeAge$) and apply the double logarithmic $Loglog$ transformation:

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

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

By definition, $RelativeAge$ is between 0 to 1 and $LoglogAge$ is positively correlated with age. The incorporation of gestation time is not essential. We simply include it to ensure

that RelativeAge takes on positive values. We used the double logarithmic transformation to link relative age to the covariates (cytosines) for the following reasons. First, the transformation maps the unit interval to the real line. Second, this transformation ascribes more influence to exceptionally high and low age values (**Extended Data Fig.1a-c**). Third, this transformation is widely used in the context of survival analysis. Fourth, this non-linear transformation worked better than the identity transformation.

The regression model underlying universal clock 2 predicts $LoglogAge$. To arrive at the DNA methylation based age estimate, one needs to apply the inverse transformation to $LoglogAge$ based on the double exponential transformation:

$$DNAmAge = exp\left(-exp\left(-LoglogAge\right)\right) * (MaxLifespan + GestatT) - GestatT \quad (3)$$

All species characteristics (e.g. MaxLifespan, 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) or mice (4 years), the same cannot be said for any other species. To address this concern, we made the following assumption: the true maximum age is 30% higher than that reported in AnAge for all species except for humans and mice (Mus musculus). Therefore, we multiplied the reported maximum lifespan of non-human or non-mouse species by 1.3. Our predictive models turn out to be highly robust with respect to this assumption.

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: it takes the logarithmic form when the

chronological age is young and takes the linear form otherwise. It is continuously differentiable at the change.

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

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

where the addition of *GestationT* ensures that the $RelativeAdultAge$ is always positive. To model a faster rate of change during development, we used a log-linear transformation on $RelativeAdultAge$ based on a function that generalizes the original transformation proposed for the human pan-tissue clock[1]:

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

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

where x denotes RelativeAdultAge. This transformation ensures continuity and smoothness at the change point at $x = m$.

In the following, we describe the estimation of the parameter m. To ensure maximum value of y to be the same across all species, the parameter m should be proportional to the maximum of x for each species, i.e. the best value for m would be the oracle value $m^* = c_1(\frac{MaxLifespan+GestationT}{ASM+GestationT})$ (**Extended Data Fig.7d**).

The proportionality constant c_1 controls the distribution of y. We chose it so that y follows approximately a normal distribution with mean zero. Since we wanted to define Clock 3 without using $MaxLifespan$, we used the ratio $\frac{GestationT}{ASM}$ to approximate the oracle value

m^* by fitting the following regression model with all the species available in our anAge database,

$$log\ \frac{MaxLifespan + GestationT}{ASM + GestationT} \approx 2.92 + 0.38 * log\ \frac{GestationT}{ASM}. (7)$$

The two log variables in the formula 7 have moderate correlation (r=0.5). Subsequently, we defined $\hat{m}$ as follows

$$\hat{m} = c_2 \left(\frac{GestationT}{ASM}\right)^{0.38}, \qquad (8)$$

where $c_2 = c_1 e^{2.92}$. We chose $c_2 = 5.0$, so that $LoglinarAge$ (termed as y in equation (5)) follows approximately a normal distribution with mean 0 (median = 9.0x10^{-4}, skewness = -0.02, **Extended Data Fig.7f**).

Setting $x = RelativeAdultAge$ in equation (5) results in

$$f(RelativeAdultAge;\ \hat{m}) = \begin{cases} \frac{RelativeAdultAge}{\hat{m}} - 1, & RelativeAdultAge \geq \hat{m} \\ \log \frac{RelativeAdultAge}{\hat{m}}, & RelativeAdultAge < \hat{m} \end{cases}. (9)$$

Universal Clock 3 predicts $LoglinearAge$ (denoted as y). To arrive at an age estimate, we use both equation 4 and equation 6 to arrive at

$$DNAmAge = \begin{cases} \hat{m} * (ASM + GestationT) * (y + 1) - GestationT & , y \geq 0 \\ \hat{m} * (ASM + GestationT) * e^{y} - GestationT & , y < 0 \end{cases}. (10)$$

Statistics for performance of model prediction

To validate our model, we used DNAm age estimates from LOFO and LOSO analyses, respectively. At each type of estimate, we performed Pearson correlation coefficients and computed median absolute error (MAE) between DNAm based and observed variables

across all samples. Correlation and MAE were also computed at species level, limited to the subgroup with n≥15 samples (within a species). We reported the medians for the correlation estimates (med.Cor) and the medians for the MAE estimates (med.MAE) across species. Analogously, we repeated the same analysis at species-tissue level, limited to the subgroup with at least 15 samples (within a species-tissue category).

For **Figure 4** we evaluated the difference (Delta.Age) between the LOSO estimate of DNAmAge and chronological age at half the maximum lifespan (0.5*$maxLifespan$). As expected, Delta.Age=LOSO DNAmAge-$(0.5 * MaxLifespan)$ is negatively correlated with species maximum lifespan.

Array Converter for Mammalian array

Since the human epidemiological cohort data were generated on a different genomic platform (Illumina 450K array), we developed an imputation scheme for converting measurements between the two platforms. The Array Converter from Human 450K Array to our Mammalian array was developed based on a study of n=141 human blood samples that were profiled using both the mammalian array and the Illumina 450k array. Next, we randomly split the data into training (80 percent) and test set (20 percent) ensuring that both data sets had a similar age distribution. In the training set, we fit penalized regressions (with elastic net penalty, alpha=0.5) for each mammalian CpG (dependent variable). To ensure that the predicted beta values lie between 0 and 1, we applied a logit transformation to each target mammalian CpG, $y = \log\left(\frac{M}{1-M}\right)$, where M denotes the beta value. As covariates, we used the subset of CpGs shared between the two array platforms. We did not consider all available covariates/CpGs for a given target mammalian CpG. Rather, we focused on a subset of CpGs located in the genomic interval

surrounding each target CpGs with bandwidth w upstream and downstream of the target CpG. We selected the bandwidth $w = 60\ Mb$ since it turned out to maximize the accuracy (R squared value) of array conversions in the test set. Finally, the sparse coefficient vectors fitted from the penalized regressions were stored as the array converters from 450K CpGs to the mammalian CpGs. For each mammalian CpG M, the imputation follows

$$\widehat{M} = \frac{2^{\hat{y}}}{1 + 2^{\hat{y}}};\ \hat{y} = X_S\widehat{\beta_S},$$

where the set S denotes the sparse subset of 450K CpGs corresponding to non-zero coefficient values. The accuracy of an imputed CpGs is guided by robust biweight midcorrelation (bicor[2]) and Pearson correlation. We found that 87% CpGs in Clock 2 and 89% CpGs in Clock 3 exceeded the correlation threshold of 0.6. In calculating the Clock estimates, we replaced the methylation levels by 0.5 for the rest of CpGs that did not satisfy the threshold.

1. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
2. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).