ORIGINAL ARTICLE

CheekAge: a next-generation buccal epigenetic aging clock associated with lifestyle and health

**Maxim N. Shokhirev · Nicole S. Torosin ·
Daniel J. Kramer · Adiv A. Johnson⬥ ·
Trinna L. Cuellar⬥**

Abstract Epigenetic aging clocks are computational models that predict age using DNA methylation information. Initially, first-generation clocks were developed to make predictions using CpGs that change with age. Over time, next-generation clocks were created using CpGs that relate to both age and health. Since existing next-generation clocks were constructed in blood, we sought to develop a next-generation clock optimized for prediction in cheek swabs, which are non-invasive and easy to collect. To do this, we collected MethylationEPIC data as well as lifestyle and health information from 8045 diverse adults. Using a novel simulated annealing approach that allowed us to incorporate lifestyle and health factors into training as well as a combination of CpG filtering, CpG clustering, and clock ensembling, we constructed CheekAge, an epigenetic aging clock that has a strong correlation with age, displays high test–retest reproducibility across replicates, and significantly associates with a plethora of lifestyle and health factors, such as BMI, smoking status, and alcohol intake. We validated CheekAge in an internal dataset and multiple publicly available datasets, including samples from patients with progeria or meningioma. In addition to exploring the underlying biology of the data and clock, we provide a free online tool that allows users to mine our methylomic data and predict epigenetic age.

Keywords Epigenetic age · Aging clock · Ensemble learning · Machine learning · Simulated annealing · Buccal

M. N. Shokhirev · N. S. Torosin · D. J. Kramer ·
A. A. Johnson (✉) · T. L. Cuellar (✉)
Tally Health, New York, NY, USA
e-mail: adiv@tallyhealth.com

T. L. Cuellar
e-mail: trinna@tallyhealth.com

Introduction

Mammalian aging is a complex, multifactorial process characterized by molecular, cellular, and organ system dysfunction. Although aging remains a poorly understood process, the field has converged on a set of 12 hallmarks that become aberrant over time and can be targeted to shorten or lengthen lifespan in model organisms [1]. However, which of these hallmarks is the most foundational has yet to be determined. Two strong contenders are genomic instability [2] and epigenetic alterations [3]. Arguing for a combination of both, a recent study created transgenic mice that repeatedly experience double-stranded DNA breaks but do not accrue mutations [4]. These non-mutagenic breaks were reported to erode the epigenome, induce an accelerated aging phenotype, and elevate epigenetic age [4]. This latter finding is interesting in

the context of some epigenetic clocks correlating with age-related health outcomes in longitudinal data and responding to interventions in clinical trials [5].

Myriad epigenetic aging clocks have been developed, and they vary in their tissue-specificity, correlation with chronological age, ability to capture health, and test–retest reliability across replicates. In terms of algorithms that incorporate information from other biomarkers, the published next-generation models GrimAge2 [6], bAge [7], and DunedinPACE [8] represent the state of the art. While innovative, these models were built using methylomic data derived from blood, which can be invasive, unpleasant, and challenging to collect in a home setting. Currently, the biohorology field [9] is lacking a published clock optimized for cheek swabs, a sample type that can be painlessly and easily collected in a variety of environments. Previous research suggests that buccal tissue is a viable sample type for epigenetic age prediction [10, 11].

To address this gap, we used an innovative computational approach in conjunction with a MethylationEPIC dataset paired with health and lifestyle questionnaire data from more than 8000 diverse adults. As presented below, the result is a unique buccal clock optimized for estimating an epigenetic age value that is associated with a plethora of lifestyle and health factors.

Methods and materials

Cohort selection and survey

We selected 10,000 volunteers from a larger cohort of over 25,000 who filled out an online questionnaire and consented to collect and send in a buccal sample. We selected volunteers with valid United States addresses while maximizing demographic diversity (chronological age, gender, and race/ethnicity) as well as various lifestyle and health (see Table S1). Of the 10,000 kits sent, 8045 samples (including 190 replicate pairs) were returned and passed all quality control checks. For each of the 8045 samples, we collected responses to 11 lifestyle- and health-related questions focused on self-rated health, self-perceived aging, sleep quality, stress levels, social satisfaction, fraction of a diet that is plant-based, exercise frequency and intensity, smoking history,

weekly alcohol consumption, relative immune health, and BMI based on self-reported weight and height. Beyond gender, we asked three demographic questions: date of birth, race/ethnicity, and education level achieved. We also predicted sample sex using methylation intensity across chromosomes and estimated cell-type proportions via the methylation data directly (described in the Supplementary Methods). To calculate correlations to lifestyle, health, and demographics, survey responses were scaled to a value between 0 and 1. Binary demographic variables were arbitrarily assigned either −1 or 1.

Sample collection

A total of 10,000 volunteers were mailed a buccal collection kit, which consisted of two VARE (Shenzhen City, Guangdong, China) flocked swabs (cat. no. VF106-80), two Mawi DNA Technologies (Pleasanton, California, USA) iSWAB-Discovery Human DNA collection devices (cat. no. ISF-T-DSC), customized instructions, and mailing pouches. Volunteers were asked to perform two collections within a 24-h period, send back both replicate samples, and register their kits. Collection instructions are provided in Supplementary Table 1.

EPIC array

Samples were preprocessed at Tempus Labs (Peachtree Corners, Georgia, USA) according to Illumina's (San Diego, California, USA) protocols for MethylationEPIC array preprocessing and loaded onto MethylationEPIC arrays. While most samples were run using the combined DNA from both collection devices to improve yields, we manually selected samples from 190 diverse individuals in our cohort to be run as replicates.

Constructing the CheekAge clock

Briefly, the EPIC arrays were preprocessed using the minfi (v 1.44.0) [12] preprocessing pipeline. CpGs were then filtered until only approximately 200,000 higher-quality CpGs remained. These higher-quality CpGs were then clustered based on their methylation pattern across the entire dataset, and the top 10,000 clusters were calculated by averaging the CpGs in each cluster, resulting in the set of independent

variables for model training. A simulated annealing [13] approach was used to train linear models optimizing for model accuracy, significance of the correlation between delta age and lifestyle/health factors, and model complexity. Finally, the model training was repeated 1098 times, and a weighted mean of the top 100 scoring models was used to predict Cheek-Age. Please see the Supplemental Methods for extensive details on preprocessing, CheekAge clock construction, clock metrics used for evaluating the clock, and the key functions and arguments used.

Evaluating CheekAge in external, publicly available datasets

Raw MethylationEPIC data was downloaded and CheekAge predictions required averaging of CpG M values for 10,000 CpG clusters. Inputs were then used to predict CheekAge using the weighted 100 ensemble model calculations. Briefly, we reprocessed 10 publicly available EPIC array datasets including an external buccal dataset (GSE111165), a blood methylation dataset looking at SARS CoV-2 infection (GSE167202), a human skin dataset from progeria patients (GSE151617), a dataset exploring accelerated aging in childhood cancer survivors (GSE197674), a meningioma dataset (GSE183647), a primary human fibroblasts dataset (GSE179847), a rectal sample dataset (GSE216024), a colorectal cancer dataset (GSE199057), a dataset looking at cultured epithelial cells antagonized with rhinovirus (GSE172365), and a melanocytic nevi dataset (GSE188593). Significance of association with delta age was calculated using linear models considering available confounding variables. We describe the specific datasets and linear model tests in the Supplemental Methods section and Supplementary Table 10.

Results

Cohort information and chronological age trends

We started by collecting buccal DNA samples and digital questionnaire responses from 8045 volunteers. Two replicate DNA samples per subject were obtained, allowing us to measure variability in methylation caused by biological and technical noise for a subset of 190 participants with sufficient DNA

collected. The 8045 EPIC samples were used in combination with the lifestyle and health information to build CheekAge, a next-generation epigenetic aging clock that correlates with chronological age, lifestyle, and health. Importantly, our cohort of 8045 samples is the largest adult buccal methylomic cohort that we are aware of, includes a chronological age range of 18 to 93 years (Fig. S1a), a similar distribution of sexes (Fig. S1b), and is diverse (Fig. S1c). Please see Supplementary Table 1 for detailed demographic information and Supplementary Table 2 for all questionnaire responses.

We started by dissecting the questionnaire responses (Table S2), finding unique patterns across chronological age (Fig. S1d, e). All variables except sex were significantly associated with chronological age (Table S3). Compared to younger respondents, older respondents tended to self-rate their health as better ($P < 2e-16$), to feel younger than their chronological age ($P = 2.68e-03$), to have lower stress levels ($P < 2e-16$), to be more socially satisfied ($P = 3.26e-07$), and to get sick less often ($P < 2e-16$). For the categories of self-rated health, self-perceived aging, sleep quality, education, and social satisfaction, we noticed that the median chronological age for the response corresponding to moderately lower values (0.25 or 0.2) was consistently the youngest than for all other responses. Older respondents typically had a more plant-based diet ($P < 2e-16$) and smoking was reported to be less common among younger respondents ($P < 2e-16$). Lastly, we saw that chronologically older respondents tended to be white while more of our chronologically younger respondents identified as non-white ($P < 2e-16$). Please see Supplementary Table 3 and 4 for additional details.

We then analyzed sex-specific differences in survey responses (Fig. S2 and Tables S3 and S4). The chronological age range for male and female respondents was the same, but the median chronological age for females was 4 years older. For males, all survey responses associated significantly with chronological age except self-perceived aging, alcohol, and education. For females, all survey responses except exercise intensity significantly associated with chronological age. Two responses showed the greatest sex-specific trends. The first was education, where education level correlated significantly with chronological age for female respondents (Ps $< 2e-16$) but not for male

respondents ($P = 0.0742$). The second was BMI, where male BMI significantly decreased with chronological age ($P = 1.83e−08$) while female BMI increased with chronological age ($P = 9.19e−04$). Interestingly, self-perceived aging and alcohol use in males were not significantly correlated with chronological age. However, self-perceived aging and alcohol use were significantly correlated for females ($P = 1.02e−04$ and $P = 1.86e−04$, respectively).

Building a next-generation buccal clock relevant to lifestyle and health

Similarly to a prior study [10], we first trained a standard penalized linear regression model using a tenfold cross validation approach. This first-generation clock correlated highly with chronological age (Fig. S3a), but delta age (epigenetic age–chronological age) failed to significantly correlate with survey factors (Fig. S3b, c). The one exception was alcohol

use, which was significantly associated with a false discovery rate (FDR) < 0.002.

To improve upon this, we created a custom objective function that included the root mean square error (RMSE) of the predicted age as well as the log of the significance of correlations between delta age and answers to survey questions. We then used a simulated annealing [13] optimization strategy along with a number of model building strategies to construct our next-generation epigenetic clock, including extensive CpG filtering to remove noisy or biased CpGs, training on clusters of CpGs to minimize noise, and using a weighted ensemble of models to further improve clock accuracy and reproducibility (Fig. 1a). As an intermediate test, we calculated the principal components (PCs) of the clustered CpG inputs and saw significant correlations between survey factors as well as technical covariates with the first 18 PCs (Fig. 1b and Table S5). All lifestyle and health factors showed strong correlations with one of the 18 PCs (Fig. 1b), and chronological age correlated strongly with PCs
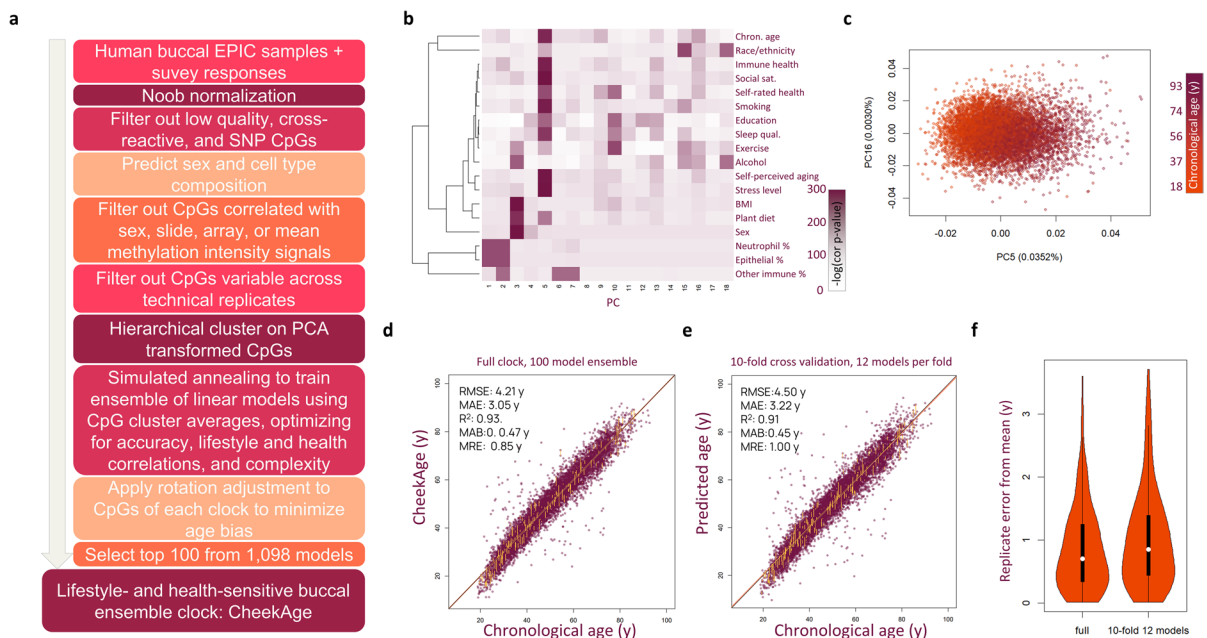
Fig. 1 Construction of the CheekAge clock. **a** Workflow for building the CheekAge clock. **b** Clustered heatmap showing correlation of lifestyle/health, demographic, and technical variables with the first 18 principal components (PCs). **c** Scatterplot showing that chronological age signals are captured in PC5 and PC16. **d** Scatterplot showing CheekAge vs chronological age trained on the entire dataset. **e** Scatterplot showing predicted age using a tenfold cross validation approach, with

12-model ensembles trained for each fold. For **d** and **e**, RMSE (root mean squared error), MAE (mean absolute error), R^2 (squared Pearson correlation), MAB (mean absolute bias), and MRE (mean replicate error from the mean) are shown. Yellow lines connect replicates taken within a 24-h period. **f** Violin plots showing the MRE for the full clock and the tenfold cross validated 12-model ensemble

5 and 16, (Fig. 1b, c). As a consequence of using a stochastic simulated annealing optimization strategy (Fig. S4a), each optimization yielded a different clock with a unique final model score. We repeated this process 1098 times for the full dataset to yield 1098 models, resulting in a distribution of final model scores, accuracies, complexities, and lifestyle correlations (Fig. S4b and Table S6). Typically, there was a tradeoff between model accuracy, complexity, and lifestyle/health correlations, with the highest-scoring and lowest-scoring models exemplifying this difference (Fig. S4c). By combining different numbers of models into an ensemble, we were able to improve accuracy, reproducibility, and correlation with survey responses at the cost of model complexity (Fig. S5). We determined that 100 models produced optimal scores, after which the improvements were marginal. Therefore, the top 100 scoring models were combined using a weighted averaging approach (Fig. S4d), to produce our final CheekAge clock. Most CpG clusters used as inputs were only utilized by a fraction of the 100 models (Fig. S4e), and each cluster included anywhere from 1 to over 150 CpGs (Fig. S4f).

To explore how the models differ from each other, we generated clustered heatmaps for all 100 models (Fig.S6). Models tended to use diverse patterns of cluster weights to predict age, and some clusters were consistently incorporated with negative or positive weights across models (Fig. S6a). Turning to the delta age predicted by CheekAge, we noticed that even

though model weights tended to be diverse between models, the ages predicted were similar to the ensemble average for any specific sample (Fig.S6b). Importantly, models with the lowest optimization score and hence highest contribution toward CheekAge were dispersed among the model clusters.

Taken together, our CheekAge clock was highly predictive of chronological age, had low age bias, and a low test–retest error when training and predicting in our full dataset (Fig. 1d) or when using a tenfold cross validation approach that uses an ensemble of 12 models per fold to estimate expected error in new similar data (Fig. 1e). In the cross-validation data, the R^2 was 0.91, the RMSE was 4.5 years, the mean absolute error (MAE) was 3.22 years, and the mean absolute bias (MAB), an indicator of chronological age bias, was 0.45 years. We observed a mean replicate error from the mean (MRE) of 0.85 years and 1 year in the full and cross-validated versions of our clock, respectively (Fig. 1f).

Since lifestyle/health responses were correlated with other lifestyle/health responses, demographics, and cell types (Fig. 2a and Table S7), we modeled delta age as a linear combination of all lifestyle/health, demographic, and technical variables to estimate the significance of each lifestyle factor correlation with all other factors held constant. Using a FDR cut-off of 0.05, we found that BMI, smoking, alcohol, social satisfaction, stress levels, exercise, sleep quality, and percent of diet that is plant-based were
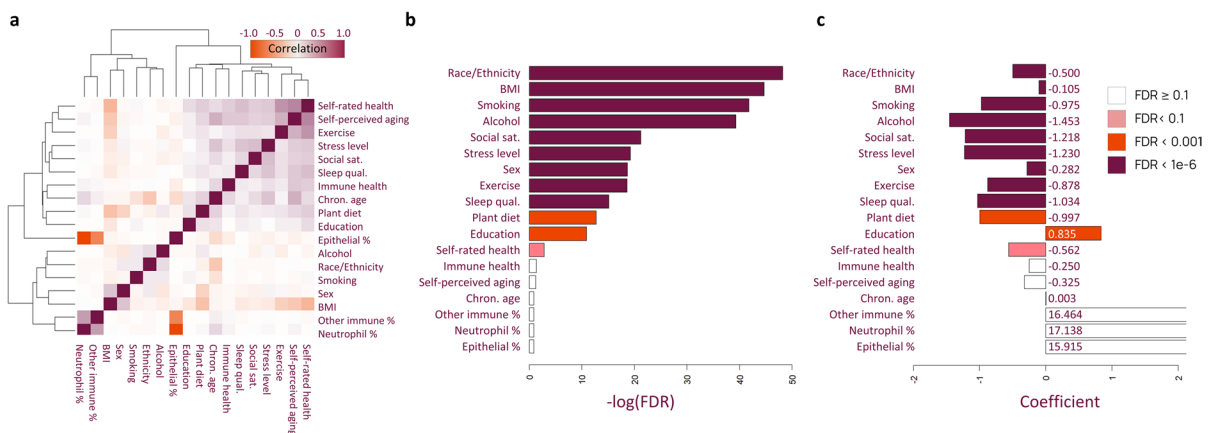
Fig. 2 Lifestyle and health factors correlate significantly with delta CheekAge. **a** Clustered heatmap showing Pearson correlation of lifestyle/health, demographic, and technical variables across all samples. **b** Significance of each variable when mode-ling delta CheekAge as a linear combination of survey factors. **c** Linear model coefficients for delta CheekAge modeled as linear combination of variables

correlated with delta age (Fig. 2b). Self-rated health displayed a FDR of 0.0586. Importantly, coefficients of the linear fit for all of these factors were changing as expected, with healthier lifestyles predicted to decrease delta age (Fig. 2c, Fig. S7, and Table S8). Lifestyle and health correlations changed depending on which model in the ensemble was used to predict age, and this was largely unaffected by the relative weight of each model (Fig.S6c). Sex-specific correlations between delta age and lifestyle/health factors are summarized in Supplementary Fig. 8 and Supplementary Table 7.

Validating the CheekAge clock

We next wanted to test how our CheekAge clock performs in other datasets (Table S9). We started by testing its performance in an independently collected buccal dataset ($n = 225$) with an age range of 18–100 years. Our clock performed well in this dataset, with a R^2 of 0.92, a MAE of 3.48 years, and a MAB of 1.19 years (Fig. 3a). We then downloaded and reprocessed a publicly available dataset [14] containing multi-tissue data from patients with intractable epilepsy. The R^2 was 0.52 and the MAE was 4.69 years in buccal tissue ($n = 16$), mostly owing to one outlier that was predicted to be much younger (Fig. 3b). The accuracy metrics were higher in saliva ($n = 15$), with a R^2 of 0.89 and a MAE of 3.59 years (Fig. 3c). In blood ($n = 15$), the R^2 was 0.82 and the MAE was 7.83 years (Fig. 3d).

We also tested whether our clock was able to pick up health- and age-related signals in other datasets [15–20]. To do this, we downloaded and reprocessed six different publicly available datasets containing health and/or disease information (Fig. 4, Fig. S9, and Table S10). As shown in Fig. 4 and Supplementary Fig. 9, we observed significant correlations between delta age and SARS-CoV-2 infection (Fig. 4a), a non-SARS-CoV-2 infection (Fig. 4a), progeria (Fig. 4b), cancer survivors who underwent abdominal/pelvic radiation treatment, alkylating agent treatment, or corticosteroid treatment (Fig. 4c), meningioma (Fig. 4d), fibroblast passaging (Fig. 4e), and BMI (Fig. 4f). We also analyzed additional datasets [21, 22] which showed harder-to-interpret associations, namely a positive correlation with tumor formation but a negative correlation with colorectal cancer, some associations in rhinovirus-antagonized epithelial cells and

steroid treatments, and an association with both navel dysplasia and control skin samples in a melanocytic nevi dataset (Table S10). Please see the Supplementary Methods for additional information, including which variables were controlled for in each analysis.

We next compared CheekAge to four other epigenetic clocks for which all CpGs were available in our dataset. Specifically, we analyzed RMSE, MAE, R^2, age bias, and test–retest error of PhenoAge [23] (Fig. 5a), Horvath et al. [24] (Fig. 5b), Zhang et al. [25] (Fig. 5c), and PedBE [11] (Fig. 5d) alongside CheekAge (Fig. 5e). As would be expected since none of these clocks were optimized for adult buccal tissue, CheekAge displayed the best overall performance (Fig. 5f), even after controlling for systematic age bias using the same rotation transformation applied to CheekAge (Fig. S10). All of the non-CheekAge clocks also showed dramatically poorer correlation between delta age with questionnaire responses, regardless of whether or not age bias rotation was applied (Fig. S11). Predicted ages and correlations of delta ages with lifestyle/health in both rotated and unrotated versions of the clocks can be found in Supplementary Table 11 and 12, respectively. Clock performance in various validation datasets, including an additional dataset [26], is summarized in Supplementary Fig. 12 and Supplementary Table 13.

Deriving biological insights from our dataset and clock

We were also interested in what insights our clock could provide into the biology of aging. To see how CpG methylation values change with chronological age in our dataset, we binned our 8045 datasets into 15 chronological age bins roughly 5 years apart. We then calculated the average methylation values of our approximately 200,000 filtered CpGs and observed four peaks (Fig. 6a). Clustering the CpGs specific to the four peaks, we noticed that there were two types of behaviors captured: CpGs that gradually increase or decrease with increasing chronological age, and those that increase or decrease only at chronological ages > 90 (Fig. 6b).

Similarly to before [27], we used a network topology–based enrichment analysis [28] to identify gene ontology (GO) [29] terms enriched among genes associated with CpGs from each of those clusters. Biological processes associated with transcription,
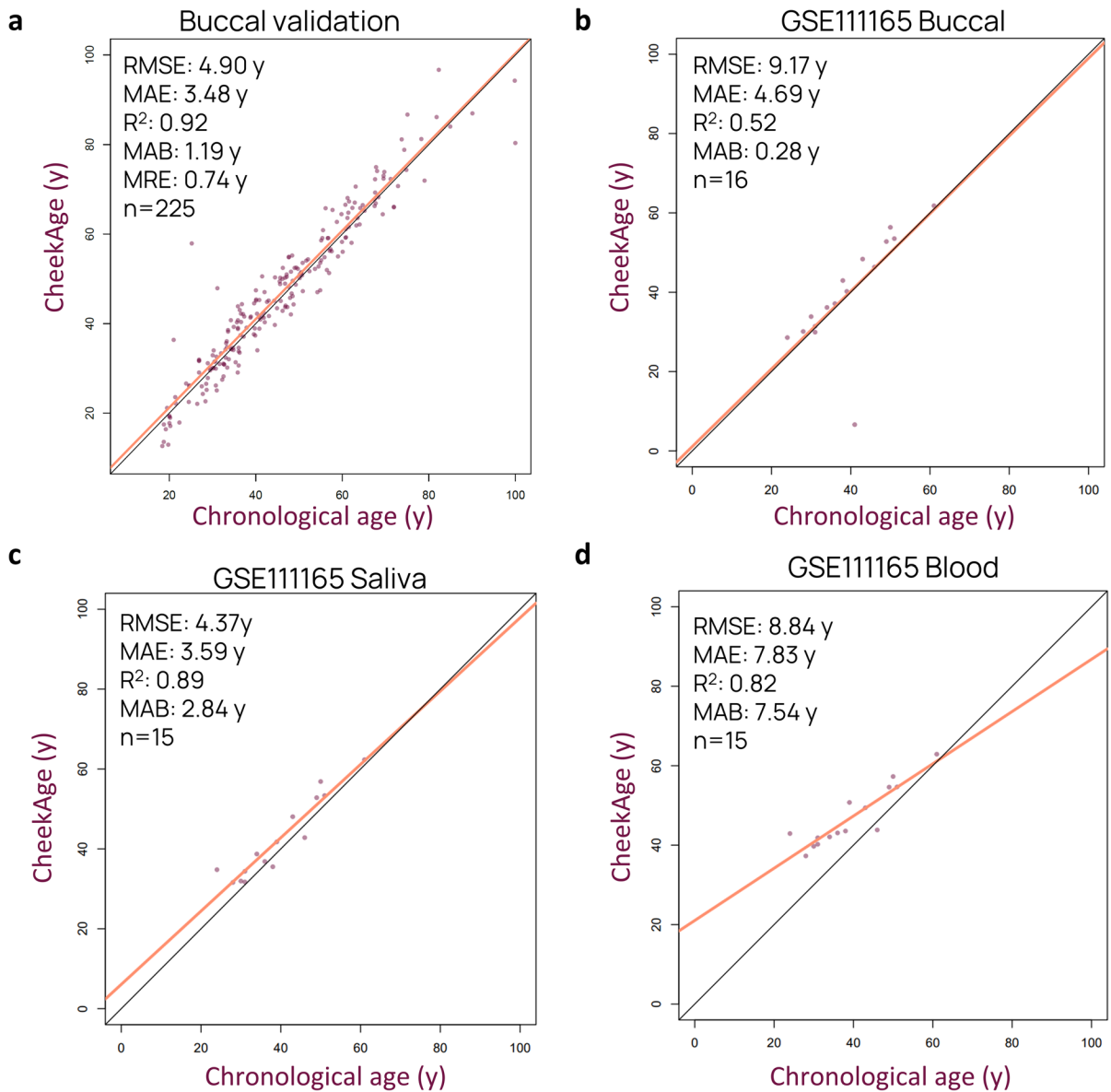
Fig. 3 Ability of CheekAge to predict age in other datasets. **a** Scatterplot and clock statistics for a buccal validation dataset that was collected the same way as the training dataset but held out during training. The chronological age range of subjects in the validation dataset is 18–100 years. **b–d** Scatterplot and clock statistics for a publicly available dataset with methylation data from tissues of medically intractable epilepsy patients. **b** Buccal, **c** saliva, and **d** blood tissue predictions are shown. The chronological age range of subjects in the publicly available dataset is 24–61 years

cell–cell signaling and hormone signaling, cell-cycle, and protein metabolism were significantly enriched among genes whose methylation gradually increases with age (Fig. 6c). Genes whose methylation dramatically jumps at very old age were enriched for cell-cycle, DNA damage response, autophagy, organelle assembly and localization, and viral process (Fig. 6d).

Processes enriched among genes associated with CpGs that decrease methylation gradually with chronological age include protein metabolic processes and hemostasis, and synapse/vesicle signaling (Fig. 6e). Finally, terms enriched among genes associated with CpGs that lost their methylation dramatically at old chronological age include cell cycle, surface receptor
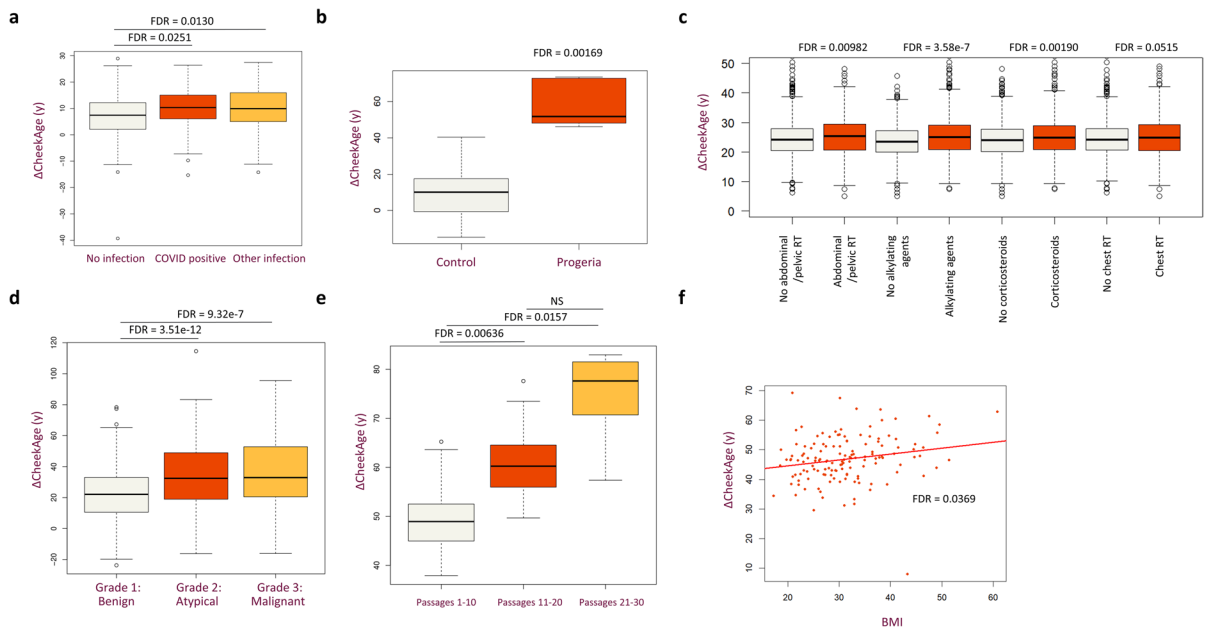
Fig. 4 Ability of CheekAge to associate with health signals in external datasets. **a** Compared to COVID-negative individuals ($n = 296$), delta age was significantly elevated in blood from COVID-positive individuals ($n = 164$) or individuals with a non-COVID acute respiratory infection ($n = 65$). **b** Relative to controls ($n = 27$), delta age was similarly significantly increased in progeria samples ($n = 9$) in a skin dataset. **c** In a blood dataset from adult survivors of childhood cancers ($n = 2138$), delta age significantly correlated with abdominal/pelvic radiation therapy, alkylating agent treatment, and corticosteroid treatment. **d** Compared to benign meningiomas ($n = 388$), meningiomas classified as atypical ($n = 142$) or malignant ($n = 35$) are predicted to be significantly older. **e** Delta age significantly increased with passage number in fibroblasts derived from healthy people. Specifically, passages 11–20 ($n = 49$) and 21–30 ($n = 18$) were significantly higher compared to passages 1–10 ($n = 51$). **f** BMI and delta age significantly correlated with one another in colorectal samples ($n = 140$)

cell–cell signaling, cellular component organization, and peptidyl-threonine modification (Fig. 6f).

We additionally explored how CpG variability changed with chronological age and delta Cheek-Age. We began by taking the same cohorts of samples binned into 5-year age bins and calculated the variance of the approximately 200,000 CpGs among samples in each of those age bins and then correlated the CpG variance with chronological age. We noticed a general increase in CpG variance with chronological age (Fig. S13a) and clustered the variance of CpGs with absolute correlation greater than 0.5 into three main clusters (Fig. S13b). We then took the genes associated with the CpGs in each of the top clusters and calculated enrichment of GO terms using a network topology–based enrichment approach. The largest cluster, which included CpGs that peaked in variance around the late 70s and mid-80s age ranges, was significantly enriched for terms associated with

cornification, keratinization, intracellular signaling, cell cycle transition, and DNA-templated transcription (Fig. S13c). Similar to the first cluster, the second largest cluster containing CpGs that peaked in variance in the very late 80s and 90s age ranges was significantly associated with DNA-template transcription, intracellular receptor signaling, cell-cycle processes, and transcription (Fig. S13d). The final cluster contained CpGs that decreased in variance with chronological age and were associated with mRNA processes, splicing, cornification, intracellular signaling, and development (Fig. S13e).

We wondered if the top-weighted clusters among the 100 models comprising our clock were enriched for specific biological processes or terms. We started by selecting the CpG clusters with absolute averaged weights greater than two, which represented the top 1.33% of all CpG clusters (Fig. S14a). The top negative weighted model
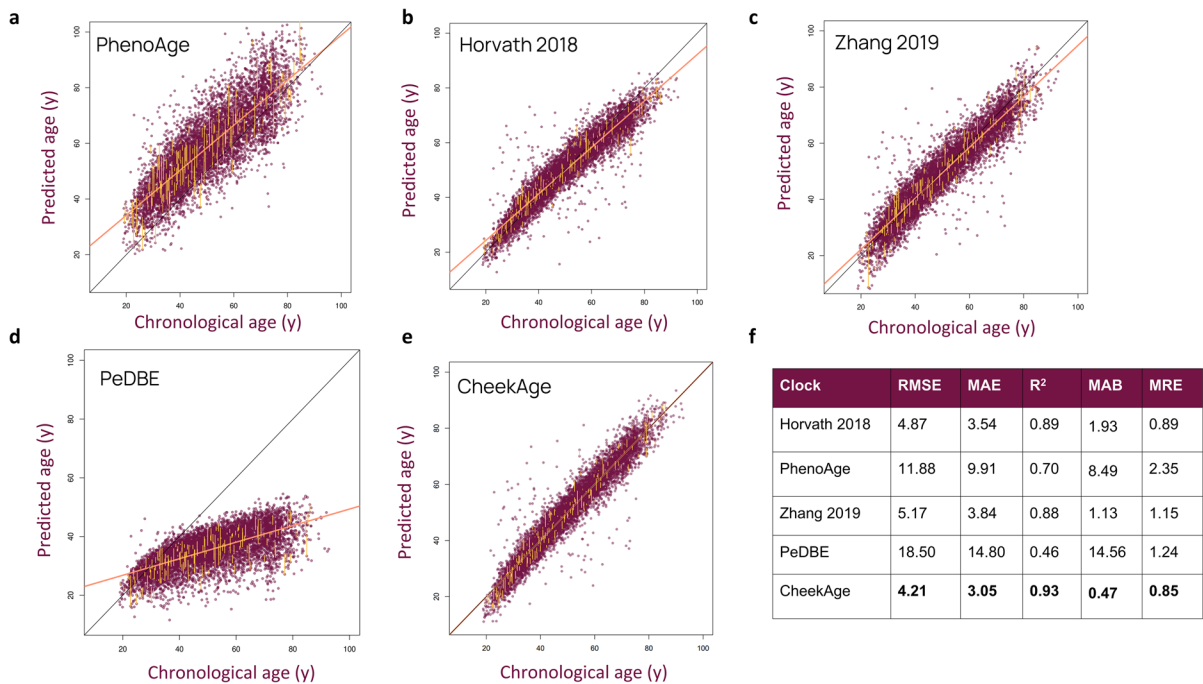
Fig. 5 Predicted age of 8045 adult buccal samples using external clocks. Panels show chronological age compared to predicted age for **a** PhenoAge, **b** Horvath et al. [24], **c** Zhang et al. [25], and **d** PedBE alongside **e** CheekAge. Replicate samples are indicated by vertical yellow lines and the linear model fit is indicated by orange diagonal lines. **f** Summary of metrics for each model fit

clusters were significantly associated with regulation of stress-activated MAPK cascade, platelet-derived growth factor signaling, and mitochondrion organization (Fig. S14b). Meanwhile, the top positive weighted model clusters were associated with various developmental and differentiation pathways, especially of the nervous system, as well as autophagy, protein localization, and cell–cell signaling (Fig. S14c).

We next explored the genomic annotation enrichment of CpGs. Taking the genomic annotations of the entire EPIC array as background, we calculated the enrichment of four sets of CpGs among CpG islands, CpG shores, CpG shelves, and Open Sea genomic elements (Fig. S14d). Overall, we noticed an enrichment of CpGs among CpG Islands, CpG shores, and CpG shelves, and a corresponding depletion of CpGs among Open Sea elements for the approximately 200,000 CpGs used to predict CheekAge, the CpGs corresponding CpG clusters with absolute weights > 2, the top 100 age-correlated CpGs, and CpGs from the top 100 age-correlated CpG clusters.

We then asked whether there was GO term enrichment among genes associated with the CpGs that were most correlated with delta CheekAge. We started by taking the CpGs with delta age correlation greater than 0.2 (0.38%) or less than −0.2 (0.14%), of the approximately 200,000 CpGs used for clock training (Fig. S15a). We found that DNA-binding transcription factor activity, neuronal development and synaptic transmission, IkB/NFkB signaling, and post-transcriptional regulation were significantly enriched among genes associated with CpGs negatively correlated with delta CheekAge (Fig. S15b). Meanwhile, nervous system development, cell–cell signaling, cell death, and MAPK cascade were significantly enriched among genes associated with CpGs positively correlated with delta CheekAge (Fig. S15c).

Next, differentially variable CpGs were identified between cohorts with relatively low (< −5 years, representing 8.5%) and relatively high (> 5 years, representing 9%) delta age values (Fig. S15d). GO terms significantly enriched among genes associated with CpGs that are more variable for lower
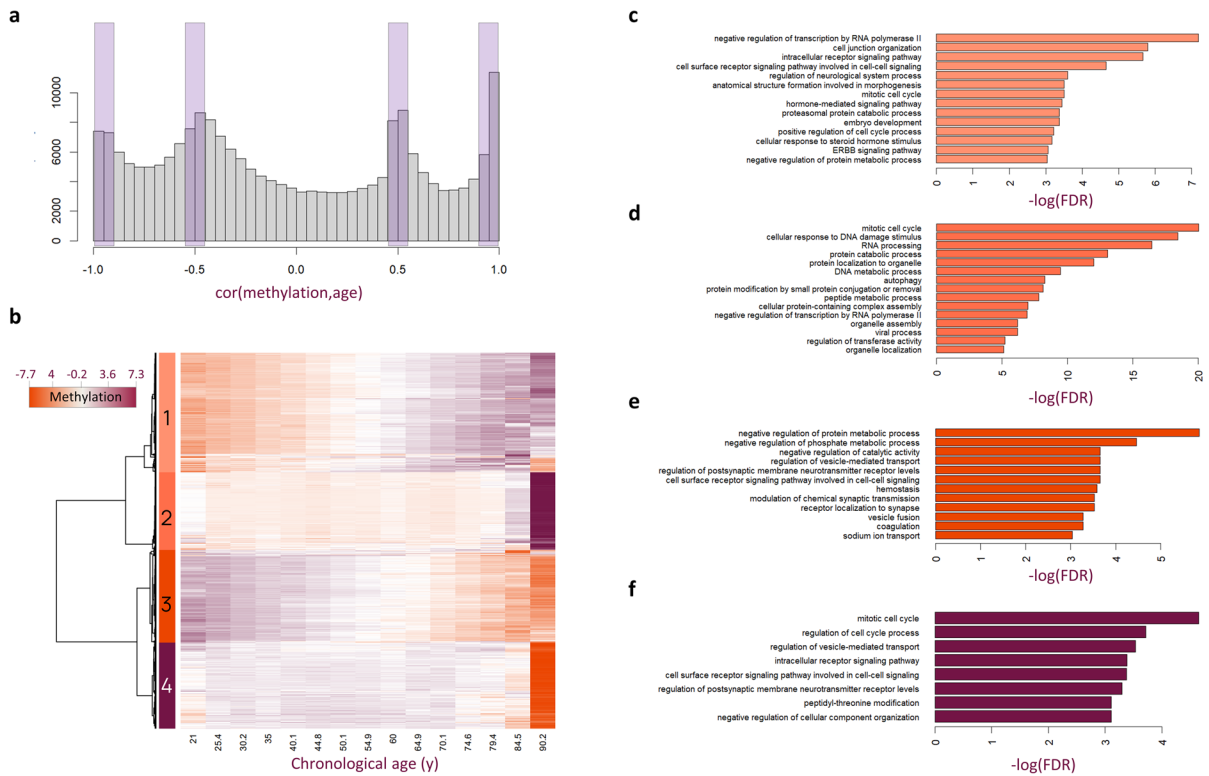
Fig. 6 Methylation patterns are correlated with age and show enrichment for biological processes. **a** Samples were binned by chronological age and the distribution of Pearson correlations of the mean methylation value of high-quality CpGs across age bins is shown. CpGs with absolute correlations near 0.5 and 1 were selected (purple rectangles) for heatmap construction. **b** Heatmap showing methylation patterns for selected highly correlated and anti-correlated CpGs with age. The top four clusters were selected for enrichment analyses. **c–f** Network topology-based enrichment analysis was performed on genes associated with cluster 1 (**c**), cluster 2 (**d**), cluster 3 (**e**), and cluster 4 (**f**) CpGs. The top 15 categories with less than 1000 genes were selected using maximum weighted set cover and plotted as bar plots of the negative log of the FDR

delta CheekAges included cell cycle and DNA damage terms, protein localization to organelles, RNA processing, and calcium ion transport (Fig. S15e). GO terms significantly enriched among genes associated with CpGs that are more variable in samples with higher delta ages involved developmental terms (Fig. S15f). The full details of the enrichment results described in this section are tabulated in Supplementary Table 14.

To explore the relationship between CpGs and lifestyle and health factors, we calculated the correlations between all of our approximately 200,000 high-quality CpGs with lifestyle/health factors, chronological age, sex, race/ethnicity, and predicted cell type compositions (Fig. S16a). We then took the top 100 most correlated or anticorrelated CpGs associated with each of the 18 factors considered (Table S15)

and calculated the overlap between them to see if each factor was associated with unique or shared CpGs (Fig. S16b). We found that cell-type proportions, alcohol, smoking, race/ethnicity, and exercise typically correlated with unique sets of top 100 CpGs, while CpGs associated with chronological age, stress, social satisfaction, education, immune health, and self-perceived aging were typically shared with other factors. As with the individual CpGs, we determined the correlation of the CpG clusters that were used as inputs for CheekAge clock construction with the 18 lifestyle/health, cell-type, and demographic factors (Fig. S17a). We then took the 100 most correlated CpG clusters for each of the 18 factors (Table S16) and plotted the overlap between them (Fig. S17b). As with the individual CpGs, cell type, race/ethnicity, smoking, and exercise tended to correlate with unique

top clusters, although typically a higher proportion of the top clusters were shared among at least two factors. Chronological age, percentage of diet that is plant-based, stress level, social satisfaction, education level, immune health, self-rated health, self-perceived aging, and sleep quality tended to correlate with CpG clusters that were common to more than one lifestyle or health factor.

Building an interactive web application for data exploration and CheekAge prediction

To help facilitate data exploration and analysis, and for others to be able to use the CheekAge clock, we created an interactive web portal called CheekAge Explorer. This tool allows users to explore the CpG methylation values and survey responses using interactive plots. Users are also able to generate plots of any variable as a function of a second. Continuous variables are shown as scatterplots with best-fit lines, categorical data (e.g., exercise activity) can be combined with continuous data using boxplots, and two categorical variables can be explored using mosaic plots [30]. Next, users can specify CpGs using gene symbols or CpG cg IDs and plot the change as a function of various factors. Finally, if users have the beta or *M*-values from EPIC V1 or V2 arrays, CheekAge Explorer can be used to calculate CheekAge. The app is available free of charge for academic use at http://cheekage.tallyhealth.com/.

Discussion

In the course of developing this next-generation epigenetic aging clock optimized for adult buccal tissue, we noted several interesting observations. For example, several lifestyle and health factors were associated with chronological age in both sexes. In male and female subjects, chronological age initially decreased with increased self-rated health at the low end of the scale. From the middle to the higher end of increasing self-rated health, the associated chronological age increased significantly. Stress levels decreased and immune health increased with chronological age in male and female subjects, as did the percentage of a diet that was plant-based. The observation that chronologically older subjects perceived themselves as being healthier, feeling less stressed,

having increased immune health, and consuming less meat could be the consequence of bias in study participation. Among older prospective subjects, healthier people may have been more interested in participating in this research. Alternatively, there could be a bias in survivorship, wherein subjects who feel less healthy, more stressed, have decreased immune health, and consume more meat have an increased rate of early mortality and fewer of those subjects survive into old age. Survivorship bias has, for example, been reported to diminish the observable relationship between age-related macular degeneration risk and smoking [31].

There were also significant differences in how chronological age was associated with smoking and BMI. Not smoking was more strongly associated with chronological age in male subjects and increased BMI moderately associated with higher chronological age in females. In contrast, increased BMI was more strongly associated with lower chronological age in males. The weaker association between increased BMI and increased chronological age in females could be due to sex-based differences in adipose tissue phenotypes [32]. The stronger association between increased BMI and decreased chronological age in males could be again due to survivorship bias, where males with increased BMI are at greater risk of mortality and, therefore, less likely to be in the older cohort [33]. Smoking and BMI were also among the factors most significantly contributing to delta age when delta age was modeled as a linear combination of survey factors. The most significant factor was race/ethnicity, which may be because non-white participants in our beta cohort were significantly less likely to drink and smoke. The five most significant lifestyle/health factors in descending order were BMI, smoking, alcohol, social satisfaction, and stress level, all of which are relevant to health and known to be associated with mortality risk.

One finding of interest was that several methylomic clusters exhibited more variability with age. These findings corroborate a report by Slieker et al., which reported that aging is synonymous with an increase in methylomic variation [34]. According to our data, variance for distinct clusters appears to ramp up around 70 years of age. Related work from the laboratory of Dr. Tony Wyss-Coray has shown that the human plasma proteome undulates with age, with noticeable peaks of differential expression occurring

at ages 34, 60, and 78 [35]. A previous meta-analysis in human transcriptomic data analogously identified a plethora of genes that were both variably and differentially expressed after age 70 [36]. While multi-omics data from the same individuals is ultimately needed to better understand how the molecular landscape becomes aberrant with age, current evidence suggests that multiple molecular systems are dysregulated in the final decades of life.

One curious finding was that a higher delta age was significantly associated with SARS-CoV-2 and non-SARS-CoV-2 respiratory infections. While the data is mixed [37], multiple reports have linked an active SARS-CoV-2 infection to an elevated epigenetic age [38, 39]. The ability of an infection to transiently impact a biomarker is not particularly surprising. For instance, an active SARS-CoV-2 infection has been connected to a reduction in grip strength [40], a higher level of C-reactive protein [41], and an increased amount of GDF15 [42]. One possible explanation for our finding is that a subset of the CpGs used in our clock are annotated to immune system genes and may be sensitive to inflammation. Indeed, one of our enrichment analyses looked at different clusters that gained or lost methylation with age. In cluster 2, which becomes hypermethylated with age, one of the top terms was "viral process." In a separate enrichment analysis looking at clusters with a delta age correlation less than -0.2, "l-kappaB kinase/NF-kappaB signaling" was the fourth most enriched term. The dysregulation of the immune system with age in vertebrate animals is well characterized. For example, a previous epigenetic and transcriptomic analysis found that innate immune pathways are commonly dysregulated across African turquoise killifish, rats, and humans [43]. Plasma proteins associated with both the adaptive and innate immune systems are also uniquely adept at predicting age in humans [44].

Additional research is warranted to better understand how our CheekAge clock behaves over time and to uncover novel factors linked to a younger or older epigenetic age. Future investigations should also examine whether health-promoting interventions—such as adopting a Mediterranean diet, increasing weekly resistance-training, or cutting out ultra-processed food—can significantly decrease CheekAge in a randomized clinical trial setting. Since aging clocks and machine learning models are highly complex, it would also be fruitful to gain a deeper understanding

of the CpG sites and CpG clusters used by our ensemble model, including their associated biology. In addition to utilizing DNA methylation, lifestyle information, and health information, future models that use artificial intelligence to incorporate additional measurement modalities may unlock greater accuracy, reliability, and utility.

Declarations

References

1. Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: an expanding universe. Cell. 2023;186:243–78. https://doi.org/10.1016/j.cell.2022.11.001.

2. Vijg J. From DNA damage to mutations: all roads lead to aging. Ageing Res Rev. 2021;68: 101316. https://doi.org/10.1016/j.arr.2021.101316.

3. de Magalhaes JP. Ageing as a software design flaw. Genome Biol. 2023;24:51. https://doi.org/10.1186/s13059-023-02888-y.

4. Yang JH, Hayano M, Griffin PT, Amorim JA, Bonkowski MS, Apostolides JK, Salfati EL, Blanchette M, Munding EM, Bhakta M, et al. Loss of epigenetic information as a cause of mammalian aging. Cell. 2023;186(305–326):e327. https://doi.org/10.1016/j.cell.2022.12.027.

5. Johnson AA, English BW, Shokhirev MN, Sinclair DA, Cuellar TL. Human age reversal: fact or fiction? Aging Cell. 2022;21:e13664. https://doi.org/10.1111/acel.13664.

6. Lu AT, Binder AM, Zhang J, Yan Q, Reiner AP, Cox SR, Corley J, Harris SE, Kuo PL, Moore AZ, et al. DNA methylation GrimAge version 2. Aging (Albany NY). 2022;14:9484–549. https://doi.org/10.18632/aging.204434.

7. Bernabeu E, McCartney DL, Gadd DA, Hillary RF, Lu AT, Murphy L, Wrobel N, Campbell A, Harris SE, Liewald D, et al. Refining epigenetic prediction of chronological and biological age. Genome Med. 2023;15:12. https://doi.org/10.1186/s13073-023-01161-y.

8. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L, Baccarelli A, Chamarti K, Gao X, Hannon E, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife*. 2022;11. https://doi.org/10.7554/eLife.73420

9. Galkin F, Mamoshina P, Aliper A, de Magalhaes JP, Gladyshev VN, Zhavoronkov A. Biohorology and biomarkers of aging: current state-of-the-art, challenges and opportunities. Ageing Res Rev. 2020;60:101050. https://doi.org/10.1016/j.arr.2020.101050.

10 Johnson AA, Torosin NS, Shokhirev MN, Cuellar TL. A set of common buccal CpGs that predict epigenetic age and associate with lifespan-regulating genes. iScience. 2022;25:105304.

11. McEwen LM, O'Donnell KJ, McGill MG, Edgar RD, Jones MJ, MacIsaac JL, Lin DTS, Ramadori K, Morin A, Gladish N, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. Proc Natl Acad Sci U S A. 2020;117:23329–35. https://doi.org/10.1073/pnas.1820843116.

12. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9. https://doi.org/10.1093/bioinformatics/btu049.

13. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science. 1983;220:671–80. https://doi.org/10.1126/science.220.4598.671.

14. Braun PR, Han S, Hing B, Nagahama Y, Gaul LN, Heinzman JT, Grossbach AJ, Close L, Dlouhy BJ, Howard MA 3rd, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. Transl Psychiatry. 2019;9:47. https://doi.org/10.1038/s41398-019-0376-y.

15. Liu Z, Meng M, Ding S, Zhou X, Feng K, Huang T, Cai YD. Identification of methylation signatures and rules for predicting the severity of SARS-CoV-2 infection with machine learning methods. Front Microbiol. 2022;13:1007295. https://doi.org/10.3389/fmicb.2022.1007295.

16. Boroni M, Zonari A, Reis de Oliveira C, Alkatib K, Ochoa Cruz EA, Brace LE, Lott de Carvalho J. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. Clin Epigenetics. 2020;12:105. https://doi.org/10.1186/s13148-020-00899-1.

17. Dong Q, Song N, Qin N, Chen C, Li Z, Sun X, Easton J, Mulder H, Plyler E, Neale G, et al. Genome-wide association studies identify novel genetic loci for epigenetic age acceleration among survivors of childhood cancer. Genome Med. 2022;14:32. https://doi.org/10.1186/s13073-022-01038-6.

18. Choudhury A, Magill ST, Eaton CD, Prager BC, Chen WC, Cady MA, Seo K, Lucas CG, Casey-Clyde TJ, Vasudevan HN, et al. Meningioma DNA methylation groups identify biological drivers and therapeutic vulnerabilities. Nat Genet. 2022;54:649–59. https://doi.org/10.1038/s41588-022-01061-8.

19. Sturm G, Karan KR, Monzel AS, Santhanam B, Taivassalo T, Bris C, Ware SA, Cross M, Towheed A, Higgins-Chen A, et al. OxPhos defects cause hypermetabolism and reduce lifespan in cells and in patients with mitochondrial diseases. Commun Biol. 2023;6:22. https://doi.org/10.1038/s42003-022-04303-x.

20. Devall MA, Sun X, Eaton S, Cooper GS, Willis JE, Weisenberger DJ, Casey G, Li L. A race-specific, DNA methylation analysis of aging in normal rectum: implications for the biology of aging and its relationship to rectal cancer. *Cancers (Basel)*. 2022;15. https://doi.org/10.3390/cancers15010045

21. Soliai MM, Kato A, Helling BA, Stanhope CT, Norton JE, Naughton KA, Klinger AI, Thompson EE, Clay SM, Kim S, et al. Multi-omics colocalization with genome-wide association studies reveals a context-specific genetic mechanism at a childhood onset asthma risk locus. Genome Med. 2021;13:157. https://doi.org/10.1186/s13073-021-00967-y.

22. Muse ME, Bergman DT, Salas LA, Tom LN, Tan JM, Laino A, Lambie D, Sturm RA, Schaider H, Soyer HP, et al. Genome-scale DNA methylation analysis identifies repeat element alterations that modulate the genomic stability of melanocytic nevi. J Invest Dermatol. 2022;142(1893–1902):e1897. https://doi.org/10.1016/j.jid.2021.11.025.

23. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10:573–91. https://doi.org/10.18632/aging.101414.

24. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, Felton S, Matsuyama M, Lowe D, Kabacik S, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies. Aging (Albany NY). 2018;10:1758–75. https://doi.org/10.18632/aging.101508.

25. Zhang Q, Vallerga CL, Walker RM, Lin T, Henders AK, Montgomery GW, He J, Fan D, Fowdar J, Kennedy M, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological aging. Genome Med. 2019;11:54. https://doi.org/10.1186/s13073-019-0667-1.

26. Nishitani S, Isozaki M, Yao A, Higashino Y, Yamauchi T, Kidoguchi M, Kawajiri S, Tsunetoshi K, Neish H, Imoto H, et al. Cross-tissue correlations of genome-wide DNA methylation in Japanese live human brain and blood, saliva, and buccal epithelial tissues. Transl Psychiatry. 2023;13:72. https://doi.org/10.1038/s41398-023-02370-0.

27. Shokhirev MN, Johnson AA. An integrative machine-learning meta-analysis of high-throughput omics data identifies age-specific hallmarks of Alzheimer's disease. Ageing Res Rev. 2022;81:101721. https://doi.org/10.1016/j.arr.2022.101721.

28. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47:W199–205. https://doi.org/10.1093/nar/gkz401.

29. Gene Ontology C, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224. https://doi.org/10.1093/genetics/iyad031

30. Belcaid M, Bergeron A, Poisson G. Mosaic graphs and comparative genomics in phage communities. J Comput Biol. 2010;17:1315–26. https://doi.org/10.1089/cmb.2010.0108.

31. McGuinness MB, Karahalios A, Kasza J, Guymer RH, Finger RP, Simpson JA. Survival bias when assessing risk factors for age-related macular degeneration: a tutorial with application to the exposure of smoking. Ophthalmic Epidemiol. 2017;24:229–38. https://doi.org/10.1080/09286586.2016.1276934.

32. Rudnicki M, Pislaru A, Rezvan O, Rullman E, Fawzy A, Nwadozi E, Roudier E, Gustafsson T, Haas TL. Transcriptomic profiling reveals sex-specific molecular signatures of adipose endothelial cells under obesogenic conditions. iScience. 2023;26:105811. https://doi.org/10.1016/j.isci.2022.105811.

33. Global BMIMC, Di Angelantonio E, Bhupathiraju ShN, Wormser D, Gao P, Kaptoge S, Berrington de Gonzalez A, Cairns BJ, Huxley R, Jackson ChL, et al. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. Lancet. 2016;388:776–86. https://doi.org/10.1016/S0140-6736(16)30175-1.

34. Slieker RC, van Iterson M, Luijk R, Beekman M, Zhernakova DV, Moed MH, Mei H, van Galen M, Deelen P, Bonder MJ, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol. 2016;17:191. https://doi.org/10.1186/s13059-016-1053-6.

35. Lehallier B, Gate D, Schaum N, Nanasi T, Lee SE, Yousef H, Moran Losada P, Berdnik D, Keller A, Verghese J, et al. Undulating changes in human plasma proteome profiles across the lifespan. Nat Med. 2019;25:1843–50. https://doi.org/10.1038/s41591-019-0673-2.

36. Shokhirev MN, Johnson AA. Modeling the human aging transcriptome across tissues, health status, and sex. Aging Cell. 2021;20:e13280. https://doi.org/10.1111/acel.13280.

37. Franzen J, Nuchtern S, Tharmapalan V, Vieri M, Nikolic M, Han Y, Balfanz P, Marx N, Dreher M, Brummendorf TH, et al. Epigenetic clocks are not accelerated in COVID-19 patients. Int J Mol Sci. 2021;22. https://doi.org/10.3390/ijms22179306

38. Cao X, Li W, Wang T, Ran D, Davalos V, Planas-Serra L, Pujol A, Esteller M, Wang X, Yu H. Accelerated biological aging in COVID-19 patients. Nat Commun. 2022;13:2135. https://doi.org/10.1038/s41467-022-29801-8.

39. Poganik JR, Zhang B, Baht GS, Tyshkovskiy A, Deik A, Kerepesi C, Yim SH, Lu AT, Haghani A, Gong T, et al. Biological age is increased by stress and restored upon recovery. Cell Metab. 2023;35(807–820):e805. https://doi.org/10.1016/j.cmet.2023.03.015.

40. Tuzun S, Keles A, Okutan D, Yildiran T, Palamar D. Assessment of musculoskeletal pain, fatigue and grip strength in hospitalized patients with COVID-19. Eur J Phys Rehabil Med. 2021;57:653–62. https://doi.org/10.23736/S1973-9087.20.06563-6.

41. Smilowitz NR, Kunichoff D, Garshick M, Shah B, Pillinger M, Hochman JS, Berger JS. C-reactive protein and clinical outcomes in patients with COVID-19. Eur Heart J. 2021;42:2270–9. https://doi.org/10.1093/eurheartj/ehaa1103.

42. Myhre PL, Prebensen C, Strand H, Roysland R, Jonassen CM, Rangberg A, Sorensen V, Sovik S, Rosjo H, Svensson M, et al. Growth differentiation factor 15 provides prognostic information superior to established cardiovascular and inflammatory biomarkers in unselected patients hospitalized with COVID-19. Circulation. 2020;142:2128–37. https://doi.org/10.1161/CIRCULATIONAHA.120.050360.

43. Benayoun BA, Pollina EA, Singh PP, Mahmoudi S, Harel I, Casey KM, Dulken BW, Kundaje A, Brunet A. Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. Genome Res. 2019;29:697–709. https://doi.org/10.1101/gr.240093.118.

44. Lehallier B, Shokhirev MN, Wyss-Coray T, Johnson AA. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. Aging Cell. 2020;19:e13256. https://doi.org/10.1111/acel.13256.