

Title: Methylation Networks Underlying Mammalian Traits

Authors: Amin Haghani^{1, 2 †*}; Caesar Z. Li^{3, 4 †}; Todd R. Robeck⁵; Joshua Zhang¹; Ake T. Lu^{1, 2}; Julia Ablaeva⁶; Victoria A. Acosta-Rodríguez⁷; Danielle M. Adams⁸; Abdulaziz N. Alagaili^{9, 10}; Javier Almunia¹¹; Ajoy Aloysius¹²; Nabil M.S. Amor¹³; Reza Ardehali¹⁴; Adriana Arneson^{15, 16}; C. Scott Baker¹⁷; Gareth Banks¹⁸; Katherine Belov¹⁹; Nigel C. Bennett²⁰; Peter Black²¹; Daniel T. Blumstein^{22, 23}; Eleanor K. Bors¹⁷; Charles E. Breeze²⁴; Robert T. Brooke²⁵; Janine L. Brown²⁶; Gerald Carter²⁷; Alex Caulton^{28, 29}; Julie M. Cavin³⁰; Lisa Chakrabarti³¹; Ioulia Chatzistamou³²; Andreas S. Chavez^{27, 33}; Hao Chen³⁴; Kaiyang Cheng³⁵; Priscila Chiavellini³⁶; Oi-Wa Choi^{37, 38}; Shannon Clarke²⁸; Joseph A. Cook³⁹; Lisa N. Cooper⁴⁰; Marie-Laurence Cossette⁴¹; Joanna Day⁴²; Joseph DeYoung^{37, 38}; Stacy DiRocco⁴³; Christopher Dold⁴⁴; Jonathan L. Dunnum³⁹; Erin E. Ehmke⁴⁵; Candice K. Emmons⁴⁶; Stephan Emmrich⁶; Ebru Erbay^{47, 48, 49}; Claire Erlacher-Reid⁴³; Chris G. Faulkes^{50, 51}; Zhe Fei^{3, 52}; Steven H. Ferguson^{53, 54}; Carrie J. Finno⁵⁵; Jennifer E. Flower⁵⁶; Jean-Michel Gaillard⁵⁷; Eva Garde⁵⁸; Livia Gerber^{59, 60}; Vadim N. Gladyshev⁶¹; Rodolfo G. Goya³⁶; Matthew J Grant⁶²; Carla B. Green⁷; M. Bradley Hanson⁴⁶; Daniel W. Hart²⁰; Martin Haulena⁶³; Kelsey Herrick⁶⁴; Andrew N. Hogan⁶⁵; Carolyn J. Hogg¹⁹; Timothy A. Hore⁶⁶; Taosheng Huang⁶⁷; Juan Carlos Izpisua Belmonte²; Anna J. Jasinska^{37, 68, 69}; Gareth Jones⁷⁰; Eve Jourdain⁷¹; Olga Kashpur⁷²; Harold Katcher⁷³; Etsuko Katsumata⁷⁴; Vimala Kaza⁷⁵; Hippokratis Kiaris⁷⁶; Michael S. Kobor⁷⁷; Pawel Kordowitzki⁷⁸; William R. Koski⁷⁹; Michael Krützen⁶⁰; Soo Bin Kwon^{16, 15}; Brenda Larison^{22, 80}; Sang-Goo Lee⁶¹; Marianne Lehmann³⁶; Jean-François Lemaître⁵⁷; Andrew J. Levine⁸¹; Xinmin Li⁸²; Cun Li^{83, 84}; Andrea R. Lim¹; David T.S. Lin⁸⁵; Dana M. Lindemann⁴³; Schuyler W. Liphardt⁸⁶; Thomas J. Little⁸⁷; Nicholas Macoretta⁶; Dewey Maddox⁸⁸; Craig O. Matkin⁸⁹; Julie A. Mattison⁹⁰; Matthew McClure⁹¹; June Mergl⁹²; Jennifer J. Meudt⁹³; Gisele A. Montano⁵; Khyobeni Mozhui⁹⁴; Jason Munshi-South⁹⁵; William J. Murphy^{96, 97}; Asieh Naderi⁷⁶; Martina Nagy⁹⁸; Pritika Narayan⁶²; Peter W. Nathanielsz^{83, 84}; Ngoc B. Nguyen¹⁴; Christof Niehrs^{99, 100}; Batsaikhan Nyamsuren¹⁰¹; Justine K. O'Brien⁴²; Perrie O'Tierney Ginn⁷²; Duncan T Odom^{102, 103}; Alexander G. Ophir¹⁰⁴; Steve Osborn¹⁰⁵; Elaine A. Ostrander⁶⁵; Kim M. Parsons⁴⁶; Kimberly C. Paul⁸¹; Amy B. Pedersen⁸⁷; Matteo Pellegrini¹⁰⁶; Katharina J. Peters^{60, 107}; Jessica L. Petersen¹⁰⁸; Darren W. Pietersen¹⁰⁹; Gabriela M. Pinho²²; Jocelyn Plassais⁶⁵; Jesse R. Poganik⁶¹; Natalia A. Prado^{110, 26}; Pradeep Reddy^{111, 2}; Benjamin Rey⁵⁷; Beate R. Ritz^{112, 113, 81}; Jooke Robbins¹¹⁴; Magdalena Rodriguez¹¹⁵; Jennifer Russell¹⁰⁵; Elena Rydkina⁶; Lindsay L. Sailer¹⁰⁴; Adam B. Salmon¹¹⁶; Akshay Sanghavi⁷³; Kyle M. Schachtschneider^{117, 118, 119}; Dennis Schmitt¹²⁰; Todd Schmitt⁶⁴; Lars Schomacher⁹⁹; Lawrence B. Schook^{117, 121}; Karen E. Sears²²; Ashley W. Seifert¹²; Aaron B.A. Shafer¹²²; Anastasia V. Shindyapina⁶¹; Melanie Simmons⁴⁵; Kavita Singh¹²³; Ishani Sinha²²; Jesse Slone⁶⁷; Russel G. Snell⁶²; Elham Soltanmohammadi⁷⁶; Matthew L. Spangler¹⁰⁸; Maria Spriggs²¹; Lydia Staggs⁴³; Nancy Stedman²¹; Karen J. Steinman¹²⁴; Donald T Stewart¹²⁵; Victoria J. Sugrue⁶⁶; Balazs Szladovits¹²⁶; Joseph S. Takahashi^{7, 127}; Masaki Takasugi⁶; Emma C. Teeling¹²⁸; Michael J. Thompson¹⁰⁶; Bill Van Bonn¹²⁹; Sonja C. Vernes^{130, 131}; Diego Villar¹³²; Harry V. Vinters¹³³; Ha Vu^{15, 16}; Mary C. Wallingford⁷²; Nan Wang^{37, 38}; Gerald S. Wilkinson⁸; Robert W. Williams¹³⁴; Qi Yan^{3, 2}; Mingjia Yao³; Brent G. Young⁵⁴; Bohan Zhang⁶¹; Zhihui Zhang⁶; Yang Zhao⁶; Peng Zhao^{14, 135}; Wanding Zhou^{136, 137}; Joseph A. Zoller³; Jason Ernst^{15, 16}; Andrei Seluanov¹³⁸; Vera Gorbunova¹³⁸; X. William Yang^{37, 38}; Ken Raj¹³⁹; Steve Horvath^{1, 2 *}

Affiliations:

- ¹ Dept. of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
- ² Altos labs, San Diego, CA, USA
- ³ Dept. of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, USA
- ⁴ Janssen Research & Development, Spring House, PA, USA
- ⁵ Zoological Operations, SeaWorld Parks and Entertainment, Orlando, Florida, USA
- ⁶ Dept. of Biology, University of Rochester, Rochester, NY, USA
- ⁷ Dept. of Neuroscience, Peter O'Donnell Jr. Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX, US
- ⁸ Dept. of Biology, University of Maryland, College Park, USA
- ⁹ Dept. of Zoology, College of Science, King Saud University, Riyadh, Saudi Arabia
- ¹⁰ Loro Parque Fundacion, Avenida Loro Parque, Puerto de la Cruz, Tenerife, Spain
- ¹¹ Dept. of Biology, University of Kentucky, Lexington, KY, USA
- ¹² Laboratory of Biodiversity, Parasitology, and Ecology, University of Tunis El Manar, Tunisia
- ¹³ Division of Cardiology, Dept. of Internal Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
- ¹⁴ Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA
- ¹⁵ Dept. of Biological Chemistry, University of California, Los Angeles, Los Angeles, California, USA
- ¹⁶ Marine Mammal Institute, Oregon State University, Newport, OR, USA
- ¹⁷ Mammalian Genetics Unit, MRC Harwell Institute, Harwell Science and Innovation Campus, Oxfordshire, UK
- ¹⁸ School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia
- ¹⁹ Dept. of Zoology and Entomology, University of Pretoria, Hatfield, South Africa
- ²⁰ Busch Gardens Tampa, Tampa, Florida, USA
- ²¹ Dept. of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, USA
- ²² The Rocky Mountain Biological Laboratory, Crested Butte, Colorado, USA
- ²³ Altius Institute for Biomedical Sciences, Seattle, WA, USA
- ²⁴ Epigenetic Clock Development Foundation, Los Angeles, CA, USA
- ²⁵ Center for Species Survival, Smithsonian Conservation Biology Institute, Front Royal, VA, USA
- ²⁶ Dept. of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA
- ²⁷ AgResearch, Invermay Agricultural Centre, Mosgiel, Otago, New Zealand
- ²⁸ Dept. of Biochemistry, University of Otago, Dunedin, Otago, New Zealand
- ²⁹ Gulf World Marine Park - Dolphin Company, Panama City Beach, FL, USA
- ³⁰ School of Veterinary Medicine and Science, University of Nottingham, UK.
- ³¹ Dept. of Pathology, Microbiology & Immunology, School of Medicine, University of South Carolina, SC, USA
- ³² Translational Data Analytics Institute, The Ohio State University, Columbus, OH, USA
- ³³ Dept. of Pharmacology, Addiction Science and Toxicology, The University of Tennessee Health Science Center, Memphis, TN, USA
- ³⁴ Medical Informatics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
- ³⁵ Biochemistry Research Institute of La Plata, Histology and Pathology, School of Medicine, University of La Plata, La Plata, Argentina
- ³⁶ Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, Dept. of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, California, USA
- ³⁷ Dept. of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA
- ³⁸ University of New Mexico, Dept. of Biology and Museum of Southwestern Biology, Albuquerque, New Mexico, United States of America
- ³⁹ Dept. of Anatomy and Neurobiology, Northeast Ohio Medical University, Rootstown, Ohio, USA
- ⁴⁰ Dept. of Environmental & Life Sciences, Trent University, Peterborough Ontario, Canada
- ⁴¹ Taronga Institute of Science and Learning, Taronga Conservation Society Australia, New South Wales, Australia
- ⁴² SeaWorld Orlando, 7007 SeaWorld Drive, Orlando, FL USA
- ⁴³ Duke Lemur Center, Durham, North Carolina, USA

44 Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, Washington, USA
 45 Altos Labs, Bay Area Institute of Science, Redwood City, USA
 46 Dept. of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA
 47 David Geffen School of Medicine, University of California, Los Angeles, CA, USA
 48 School of Biological and Chemical Sciences, Queen Mary University of London, London, UK
 49 School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK
 50 Dept. of Statistics, University of California Riverside, CA, USA
 51 Dept. of Biological Sciences, University of Manitoba, Winnipeg, Canada
 52 Fisheries and Oceans Canada, University Crescent, Winnipeg, Canada
 53 Dept. of Population Health and Reproduction, University of California, Davis School of Veterinary Medicine, Davis, CA, USA
 54 Mystic Aquarium, Mystic, Connecticut, USA
 55 University of Lyon, CNRS, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France
 56 Greenland Institute of Natural Resources, Nuuk, Greenland
 57 School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia
 58 Evolutionary Genetics Group, Dept. of Evolutionary Anthropology, University of Zurich, Zurich, Switzerland
 59 Division of Genetics, Dept. of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
 60 Applied Translational Genetics Group, School of Biological Sciences, Centre for Brain Research, The University of Auckland, Auckland, New Zealand
 61 Vancouver Aquarium, Vancouver, Canada
 62 SeaWorld San Diego, San Diego, CA, USA
 63 Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
 64 Dept. of Anatomy, University of Otago, Dunedin, New Zealand
 65 Division of Human Genetics, Dept. of Pediatrics, University at Buffalo, New York, USA
 66 Division of Infectious Diseases, Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States
 67 Department of Molecular Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
 68 School of Biological Sciences, University of Bristol, Bristol, UK
 69 Norwegian Orca Survey, Andenes, Norway
 70 Mother Infant Research Institute, Tufts Medical Center, Boston, MA, USA
 71 Yuvan Research, Mountain View, CA, USA
 72 Kamogawa Sea World, Kamogawa, Chiba, Japan
 73 Peromyscus Genetic Stock Center, University of South Carolina, SC, USA
 74 Dept. of Drug Discovery and Biomedical Sciences, College of Pharmacy, University of South Carolina, SC, USA
 75 Edwin S.H. Leong Healthy Aging Program, Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, Canada
 76 Institute of Animal Reproduction and Food Research of Polish Academy of Sciences, Olsztyn, Poland
 77 LGL Limited, King City, ON, Canada
 78 Evolutionary Genetics Group, Dept. of Anthropology, University of Zurich, Zurich, Switzerland
 79 Center for Tropical Research, Institute for the Environment and Sustainability, UCLA, Los Angeles, CA, USA
 80 Dept. of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
 81 Technology Center for Genomics and Bioinformatics, Dept. of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA, USA
 82 Texas Pregnancy and Life-course Health Center, Southwest National Primate Research Center, San Antonio, Texas, USA
 83 Dept. of Animal Science, College of Agriculture and Natural Resources, Laramie, Wyoming, USA
 84 Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, Canada
 85 Biology Department, University of New Mexico, Albuquerque, NM, USA
 86 Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

- 87 White Oak Conservation Center, Yulee FL, USA
- 88 North Gulf Oceanic Society, Homer, Alaska, USA
- 89 Translational Gerontology Branch, National Institute on Aging Intramural Research Program, National Institutes of Health, Baltimore, MD, USA
- 90 ABS Global, DeForest, WI, USA
- 91 Marineland of Canada, Niagara Falls, Ontario, Canada
- 92 Biomedical and Genomic Research Group, Dept. of Animal and Dairy Sciences, University of Wisconsin Madison, Madison, Wisconsin, US
- 93 Dept. of Preventive Medicine, University of Tennessee Health Science Center, College of Medicine, Memphis, TN, USA
- 94 Louis Calder Center - Biological Field Station, Dept. of Biological Sciences, Fordham University, Armonk, NY USA.
- 95 Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas, USA
- 96 Interdisciplinary Program in Genetics and Genomics, Texas A&M University, College Station, Texas, USA
- 97 Museum für Naturkunde, Leibniz-Institute for Evolution and Biodiversity Science, Berlin, Germany
- 98 Institute of Molecular Biology (IMB), Mainz, Germany
- 99 Division of Molecular Embryology, DKFZ-ZMBH Alliance, Heidelberg, Germany
- 100 Mammalogical Society of Mongolia, Ulaanbaatar, Mongolia
- 101 Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge, UK
- 102 Deutsches Krebsforschungszentrum, Division of Regulatory Genomics and Cancer Evolution, Heidelberg, Germany
- 103 Dept. of Psychology, Cornell University, Ithaca, NY, USA
- 104 SeaWorld San Antonio, San Antonio, Texas, USA
- 105 Dept. Molecular Cell and Developmental Biology, University of California Los Angeles, Los Angeles, CA, USA
- 106 School of Earth, Atmospheric and Life Sciences, University of Wollongong, Wollongong, Australia
- 107 Dept. of Animal Science, University of Nebraska, Lincoln, USA
- 108 Mammal Research Institute, Dept. of Zoology and Entomology, University of Pretoria, Hatfield, South Africa
- 109 Dept. of Biology, College of Arts and Science, Adelphi University, One South Avenue, Garden City, NY, USA
- 110 Salk Institute for Biological Studies, La Jolla, CA USA
- 111 Dept. of Epidemiology, UCLA Fielding School of Public Health, Los Angeles, CA, USA
- 112 Dept. of Environmental Health Sciences, UCLA Fielding School of Public Health, Los Angeles, CA, USA
- 113 Center for Coastal Studies, Provincetown, MA, USA
- 114 Miami Seaquarium, Miami, Florida, USA
- 115 The Sam and Ann Barshop Institute for Longevity and Aging Studies, and Dept. of Molecular Medicine, UT Health San Antonio, and the Geriatric Research Education and Clinical Center, South Texas Veterans Healthcare System, San Antonio TX, USA
- 116 Dept. of Radiology, University of Illinois at Chicago, Chicago, IL, USA
- 117 Dept. of Biochemistry and Molecular Genetics, University of Illinois at Chicago, Chicago, IL, USA
- 118 National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA
- 119 College of Agriculture, Missouri State University, Springfield MO, USA
- 120 Dept. of Animal Sciences, University of Illinois at Urbana-Champaign, USA
- 121 Dept. of Forensic Science, Environmental & Life Sciences, Trent University, Peterborough Ontario, Canada
- 122 Shobhaben Pratapbhai Patel School of Pharmacy & Technology Management, SVKM'S NMIMS University, Mumbai, India
- 123 Species Preservation Laboratory, SeaWorld San Diego, California, USA
- 124 Biology Department, Acadia University, Wolfville, Nova Scotia, Canada
- 125 Dept. of Pathobiology and Population Sciences, Royal Veterinary College, Hatfield, UK
- 126 Howard Hughes Medical Institute, Dept. of Neuroscience, University of Texas Southwestern Medical Center, Dallas, Texas, USA
- 127 School of Biology and Environmental Science, University College Dublin, Belfield, Dublin, Ireland
- 128 Animal Care and Science Division, John G. Shedd Aquarium, Chicago, Illinois, USA
- 129 School of Biology, The University of St Andrews, Fife, UK
- 130 Neurogenetics of Vocal Communication Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

- ¹³¹ Blizard Institute, Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK
¹³² Dept. of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, USA
¹³³ Dept. of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, College of Medicine, Memphis, TN, USA
¹³⁴ Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, USA
¹³⁵ Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, USA
¹³⁶ Dept. of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, USA
¹³⁷ Depts. of Biology and Medicine, University of Rochester, Rochester, NY, USA
¹³⁸ Altos Labs, Cambridge, UK

† Co-first authors

* Corresponding authors: shorvath@mednet.ucla.edu ahaghani@altoslabs.com

Address: 5510 Morehouse Dr San Diego, CA 92121 United States

Abstract: Using DNA methylation profiles (n=15,456) from 348 mammalian species, we constructed *phyloepigenetic* trees that bear remarkable similarities to traditional phylogenetic ones. Using unsupervised clustering across all samples, we identified 55 distinct cytosine modules, of which 30 are related to traits such as maximum lifespan, adult weight, age, sex, and human mortality risk. Maximum lifespan is associated with methylation levels in *HOX* genes, developmental processes, and potentially regulated by pluripotency transcription factors. The methylation state of some modules responds to perturbations such as caloric restriction, ablation of growth hormone receptors, consumption of high-fat diets, and expression of Yamanaka factors. This study reveals an intertwined evolution of the genome and epigenome that mediates the biological characteristics and traits of different mammalian species.

One-Sentence Summary: Phyloepigenetic trees, derived from DNA methylation profiles, mirror mammalian evolution, and are related to mammalian lifespan and other species characteristics.

Main Text:

Comparative epigenomics is a burgeoning field that integrates epigenetic signatures with phylogenetic relationships to decipher gene-to-trait functions {Xiao, 2012 #307; Villar, 2015 #306; Qu, 2018 #436}. Prior research has investigated the capacity of DNA methylation patterns in regulatory sequences to reflect evolutionary relationships among species {Qu, 2018 #436; Klughammer, 2023 #822}. A recent study compared methylation data across multiple animal species at orthologous gene promoters using a sequencing-based assay that did not specifically target conserved CpGs {Klughammer, 2023 #822}. Previous investigations faced limitations regarding the measurement platform, particularly the low sequencing depth at conserved CpGs and the sample size per species.

Our study overcomes these constraints in several ways. First, we utilized a measurement platform ensuring high sequencing depth at conserved CpGs, allowing for a more precise analysis of DNA methylation patterns in highly conserved DNA regions. Second, we increased the sample size per species, aiming for around 30 samples. We profiled 348 species from 25 of the 26 mammalian taxonomic orders. This comprehensive dataset enables examination of phylogenetic relationships,

co-methylation relationships between cytosines, and their associations with maximum lifespan and other species characteristics.

We profiled 15,456 samples (**Fig. 1A; table S1**) using a methylation array platform that provides effective sequencing depth at highly conserved CpGs across mammalian species {Arneson, 2022 #722}. This dataset is the product of the multi-national Mammalian Methylation Consortium. In previous studies, we applied supervised machine learning methods to generate DNA methylation-based predictors of age, called epigenetic clocks, for numerous species {Parsons, 2023 #830; Lu, 2021 #744; Kordowitzki, 2021 #426; Prado, 2021 #743; Robeck, 2021 #740; Larison, 2021 #747; Mozhui, 2022 #813; Sugrue, 2021 #558; Robeck, 2021 #741; Horvath, 2022 #719; Horvath, 2022 #800; Horvath, 2021 #751; Horvath, 2021 #750; Jasinska, 2021 #749; Raj, 2021 #742; Schachtschneider, 2021 #739; Cossette, 2023 #799; Lemaitre, 2022 #746; Horvath, 2022 #721; Wilkinson, 2021 #736; Wilkinson, 2021 #737; Horvath, 2022 #720; Pinho, 2022 #728; Horvath, 2022 #725; Peters, 2022 #815; Chiavellini, 2022 #816}.

Here, we perform a large-scale cross-species unsupervised analysis of the entire dataset to reveal the relationship of DNA methylation (DNAm) with mammalian phylogeny. We show that we can construct *phyloepigenetic* trees that parallel traditional phylogenetic ones. We then proceed to interrogate the extent to which DNA methylation underpins specific biological traits by employing unsupervised weighted correlation network analysis (WGCNA) to minimize the influence of bias on our observations. This approach identifies clusters of co-methylated CpGs (co-methylation modules) that are associated with species characteristics, including taxonomy, tissue type, sex, lifespan, and aging.

Results:

Evolution and DNA methylation

We generated a dataset consisting of DNA methylation profiles of 15,456 DNA samples derived from 70 tissue types, from 348 mammalian species using the mammalian methylation array {Arneson, 2022 #722}. We evaluated whether methylation levels of cytosines (CpGs) in DNA sequences that are conserved across species would allow us to construct what could be termed a phyloepigenetic tree. To avoid potential confounding by different tissue types, we generated tissue-specific phyloepigenetic trees (**Fig. 1B; fig. S2; fig. S3**). We defined the ‘*Congruence*’ between traditional phylogenetic trees and phyloepigenetic trees as the Pearson correlation coefficient between distances (branch length) based on phyloepigenetic trees and evolutionary distances in traditional phylogenetic trees. We observe high congruence (*Congruence*=0.93, **Fig. 1C; fig. S2**) for the blood-based phyloepigenetic tree (124 species), and lower congruence values for non-blood tissues (*Congruence*=0.58 for liver and *Congruence*=0.72 for skin, **fig. S2**). The lower congruence in liver (158 species) and skin (133 species) may be due to potential variability in sampling between species. The varying congruence across tissue types shows that the CpG probes do not serve as genotyping proxies. The tissue dependence of congruence indicates that phyloepigenetic trees are derived based on differences in methylation levels and not sequence conservation. This point is also corroborated by three sensitivity analyses, which confirmed that the high congruence was indeed due to differences in methylation levels (supplementary text). In particular, the phyloepigenetic trees based on the 180 CpGs with the most significant detection p values across all 348 species still are congruent with traditional trees (**fig. S2F-G**).

In order to identify CpGs that exhibit a pronounced phylogenetic signal in relation to methylation and phylogenetic trees, we utilized Blomberg's K statistic {Blomberg, 2003 #826}. Among the top 500 CpGs showing significant phylogenetic signals (nominal Blomberg $p < 0.001$, additionally selected by variance z-score), we observed an enrichment in upstream intergenic regions (odds ratio OR = 1.4, Fisher exact $p < 0.05$, **fig. S4B**). To further investigate regions with the strongest phylogenetic signal, we divided the data into groups of 10 CpGs relative to the transcriptional start site (TSS). This analysis also confirmed that intergenic regions exhibit significant phylogenetic signals (OR > 3, Fisher exact $p < 0.05$), while the promoter regions did not show such signals (**Fig. 1D**).

DNA methylation networks relate to individual and species traits

We used signed weighted correlation network analysis (WGCNA, an unsupervised analysis) {Langfelder, 2008 #31} to cluster CpGs with similar methylation dynamics across samples into co-methylation modules. We then summarized their methylation profiles as "module eigengenes". The respective eigengenes of these modules were used to identify their potential correlations with various traits within and across mammalian species.

Our data analysis proceeded in two sequential phases. First, we developed several co-methylation networks using data from 11,099 DNA samples from 174 species (discovery dataset, finalized March 2021). A eutherian network (Net1) was formed from 14,705 conserved CpGs using this dataset (**Fig. 2A**). Later (March 2022), we generated a second data set of 4,357 samples from 30 tissues of 240 mammalian species (174 new species, and 66 that are represented in the discovery set), which were not used to define modules and were used as an independent validation set. All the eutherian modules were present in the independent validation dataset according to module preservation statistics (*corKME*) {Langfelder, 2011 #586}, validating the presence of these modules (*corKME* > 0.43, $p < 10^{-22}$; median *corKME* = 0.84) (**fig. S5**). These modules were designated with colors according to the WGCNA convention (**Fig. 2A**). The smallest module (lavenderblush3) consisted of 33 CpGs, while the largest (turquoise) had 1,864 CpGs.

To characterize the 55 modules with respect to species characteristics (e.g., maximum lifespan and average adult weight), module eigengenes were calculated in all samples (discovery and validation set combined, 331 eutherian species). As information on taxonomic order, tissues, maximum lifespan, age, sex and adult weight of each species were available, we were able to assess whether any of the module eigengenes correlated with these traits. Of the 55 modules, 30 were found to be correlated with at least one trait (**Fig. 2B; fig. S7; table S3**). Specifically, 15 modules were related to taxonomic orders such as primates, rodents, or carnivores (**Fig. 2B; see also fig. S11**). Ten modules related to tissue type (**fig. S11**), while two were related to sex (**fig. S11**), one to age, seven modules to maximum lifespan, and four to average adult species weight. Some modules were related to multiple characteristics. In the following sections, we mainly focus on the modules that relate to mammalian maximum lifespan, adult weight, and age. Other modules (related to taxonomic order, tissue type, or sex) are described in the supplement (**fig. S11**). We performed two analyses to ascertain whether these eutherian modules are also applicable to marsupials and monotremes. First, we trained a network (Net2) in both eutherians and marsupials based on only 7,956 probes that are mappable to both. The color bands underneath the hierarchical tree reveal that all the Net1 modules were also preserved in Net2 (**Fig. 2A**). Second, we selected CpGs in Net1 modules that are also mapped to marsupials or monotremes and confirmed that their eigengene relationships to primary traits were retained in these mammalian clades (**table S3**). For example, the magenta module, which is related to blood in eutherians, was also found to be so in

monotremes (**table S3**), which confirms that the Net1 modules can indeed be applied to other mammalian clades, by selecting probes that are also mapped to those clades.

A functional enrichment study, accounting for the mammalian array background, revealed that the 500 most interconnected CpGs per module neighboring genes are implicated in many biological processes including development, immune function, metabolism, reproduction, stem cell biology, stress responses, aging, and various signaling pathways (**Fig. 2C, fig. S9**).

Relationship with protein-protein interactions

We examined whether the proteins encoded by cognate genes (closest to respective CpGs) within modules are known to mutually interact or predicted to do so by STRING protein-protein interaction networks, which integrate known and predicted protein associations from over 14,000 organisms {Szklarczyk, 2021 #595}. A permutation test analysis evaluating the global cluster coefficient {Barrat, 2004 #819} of each module showed that 14 modules are significantly enriched ($p < 0.001$) for genes encoding mutually interacting proteins (**Fig. 2D**). Overall, these results suggest that co-methylation relationships can be reflected at the protein level for a subset of modules.

Modules related to maximum lifespan

To adjust for potential confounders, we used four regression modeling approaches to identify modules that are associated with log transformed maximum lifespan (dependent variable): 1) a univariate regression model whose covariate was the module eigengene (averaged per species), 2) a phylogenetic regression model whose covariate was again the module eigengene (averaged per species), 3) a multivariate linear regression model that included the module eigengene, sex, tissue, and relative age as covariates, 4) model approach 1 applied to specific tissue types.

The marginal analysis identified four modules (magenta, black, midnightblue, and tan) that related significantly to maximum lifespan (the absolute value of the Pearson correlation exceeded $r=0.6$, Student T test $p < 1 \times 10^{-33}$). The CpGs underlying the implicated modules exhibit the sample patterns as can be seen from corresponding heatmaps (**fig. S14C**). Phylogenetic regression also identified associations of the same modules (**table S3**). Our fourth modeling approach, i.e. the tissue-stratified marginal analysis, indicates that the relationship of modules to maximum lifespan is often tissue-specific. For example, the magenta and midnightblue modules relate to maximum lifespan in lung and liver (**fig. S14A**). In contrast, the black module relates to maximum lifespan only in skin, while the tan module exhibited a weak relationship to lifespan in the tissue-specific analysis.

For ease of comprehension, modules were labeled with the trait and direction of relationship by superscript +/- signs (e.g. magenta = Lifespan⁽⁺⁾Weight⁽⁺⁾Blood⁽⁺⁾ module). The two modules (magenta with 480 CpGs, and midnightblue with 249 CpGs) that correlated with lifespan in lung and liver also correlated significantly with average adult weight across all eutherian species ($r=0.47$ to 0.55 , $p < 1 \times 10^{-18}$, **Fig. 3**). The magenta module (Lifespan⁽⁺⁾Weight⁽⁺⁾Blood⁽⁺⁾) is enriched with developmental genes such as *HOXA5* and *VEGFA*, *SOX2*, and *WNT11* (**table S4**). The midnightblue (Lifespan⁽⁺⁾Weight⁽⁺⁾) module implicates genes involved in tRNA metabolism ($p=2 \times 10^{-6}$, e.g. *URM1*), lipopolysaccharides ($p=5 \times 10^{-6}$, e.g. *CERCAM*), development ($p=10^{-4}$, *HOXL* gene family), and fatty acids ($p=2 \times 10^{-3}$, e.g. *ACADVL*). The magenta module also relates to lifespan and average weight of dog breeds (**Fig. 3C**, $r = -0.30$, $p=0.003$). Furthermore, it is related

to the hazard of human death (hazard ratio HR= 0.91, MetaP=0.0016, **Fig. 3D**) in epidemiological cohort studies.

After adjustment for phylogeny, the cyan module relates to mammalian lifespan phylogenetic contrast ($r=0.42$, $p=4\times 10^{-14}$, **fig. S13I**). The Lifespan⁽⁺⁾Liver⁽⁻⁾ (cyan) module consists of genes that play a role in adaptive immunity ($p=2\times 10^{-6}$), histone and protein demethylation ($p=0.0001$), and metabolism ($p=0.0004$) (**table S4**).

The multivariate model analysis included sex, tissue type, and relative age as covariates to reveal the modules that relate to lifespan in different tissues. The regression analysis found two modules with opposing correlations with maximum lifespan: green module (lifespan $r=0.42$, average weight $r=0.38$, $p<10^{-300}$) and the greenyellow module (lifespan $r=-0.44$, average weight $r=-0.35$, $p<10^{-300}$, **fig. S13J**). The CpGs of the Lifespan⁽⁻⁾Weight⁽⁻⁾Rodentia⁽⁺⁾ (greenyellow) are located near genes that play a role in development ($p=5\times 10^{-13}$, **table S4**) and in RNA metabolism ($p=6\times 10^{-12}$).

Age-related consensus module in mammals

The purple module (denoted subsequently as RelativeAge⁽⁺⁾ module) exhibited the strongest positive correlation with relative age (Relative age $r=0.35$, $p<10^{-300}$, **Fig. 3E**; **fig. S13**).

To remove the confounding effects of species and/or tissue type, we also constructed seven consensus networks (denoted cNet3,...,cNet9, description in supplement and methods). The RelativeAge⁽⁺⁾ module was preserved in 3 different consensus networks (cNet3, cNet4, and cNet6, **Fig. 2A**), suggesting conservation in different species and tissues (scatter plot in **fig. S11H**). The purple RelativeAge⁽⁺⁾ module is positively enriched for CpGs in regulatory regions (e.g. promoters and 5'UTR) and depleted in intron regions (**fig. S15**). Functional enrichment of this module highlighted embryonic stem cell regulation, axonal fasciculation, angiogenesis, and diabetes-related pathways (**table S3**). The CpGs in this module are adjacent to Polycomb repressor complex 2 (PRC2, EED) targets which are marked by H3K27me3 (**table S3**).

Ingenuity pathway analysis implicates POU5F1 (alias OCT4), SHH, ASCL1, SOX2, and NEUROG2 proteins as putative upstream regulators of the RelativeAge⁽⁺⁾ module. We used GTEx data to examine if the mRNA levels of any of these upstream regulators are altered with age in several human tissues. *OCT4* (repeated measures correlation, $rmCor=0.07$, $p=2\times 10^{-14}$), which is among the four known Yamanaka factors for cellular dedifferentiation, showed a positive increase with age in several but not all human tissues (**fig. S11F**). Nine other genes (e.g. *HOXD10*, $rmCor=0.16$, $p=4\times 10^{-50}$; *SRXN1*, $rmCor=-0.14$, $p=4\times 10^{-52}$) from the RelativeAge⁽⁺⁾ module also had a nominally significant $rmCorr$ ($p<0.005$) in GTEx data (**Fig. 3F**; **fig. S11G**), although opposite aging patterns could be found in select tissues. These observations highlight the relevance of genes in the RelativeAge⁽⁺⁾ module to stem cell biology and aging in human tissues.

Interventional studies in mice

We related our methylation modules to interventions that are known to modulate the lifespan of mice (**Fig. 4A–C**). This included growth hormone receptor knockout (i.e., dwarf mice) {Pilcher, 2003 #301} and caloric restriction {Acosta-Rodriguez, 2022 #829}, which extended life, and high-fat diet, which elicited the opposite effect {Mozhui, 2022 #813}. Six modules, including the purple module (RelativeAge⁽⁺⁾) showed a significant decrease ($p<0.05$) of the module eigengene in dwarf mice and after caloric restriction, and conversely a modest increase after a high-fat diet. While the magenta, black, midnight blue, tan, and greenyellow modules have connections to the maximum lifespan in mammals, they did not present a clear relationship with interventions that modify

murine lifespan (GHRKO, CR, and high-fat diet). This suggests a mutual exclusivity between the modules related to maximum mammalian lifespan and those affected by interventions modulating murine lifespan.

Transient expression of Yamanaka factors

We examined if a transient expression of the Yamanaka factors in the 4-factor (4F) mouse affects the module eigengenes. The experimental design is shown in **Fig. 4D**, with additional details reported in the original article {Browder, 2022 #734}. Four out of six of the above-mentioned murine intervention modules showed a nominally significant dose-dependent rejuvenation in murine skin ($p < 0.06$) and 2 modules showed the same in kidney (dose refers to the duration of 4F treatment: 0-, 1-, 7-, and 10-months intermittent expression of 4F factors) (**Fig. 4E**). The purple, ivory, and lavenderblush3 modules were particularly sensitive to the 4F treatment (Pearson correlation ≤ -0.64 in skin). In addition, the purple RelativeAge⁽⁺⁾ module's response to the 4F treatment is consistent with bioinformatic findings that OCT4 is an upstream regulator of this module. Among the lifespan modules, only the black module demonstrates an increase ($p = 0.007$) in skin of 4F treated mice, but this is not observed in the kidney.

Epigenome-wide association analysis of maximum lifespan

We carried out epigenome-wide association studies (EWAS) to identify individual CpGs with methylation levels that correlate with maximum lifespan. To reduce bias resulting from different levels of sequence conservation, our EWAS of maximum lifespan focused on 333 eutherian species, excluding marsupial and monotreme species. We restricted the analysis to 28,318 high quality probes that are conserved between humans and mice.

When relating individual CpGs to log-transformed maximum lifespan, we used several modeling approaches (detailed in the Supplementary Text). Briefly, our first approach, generic modeling, applied regression analysis ignoring tissue type and age. Second, we repeated the regression analysis after focusing on a given tissue type. Third, we focused on specific non-overlapping age groups: young animals (defined as age younger than 1.5 times the age at sexual maturity), middle-aged, and old (defined as Age > 3.5 times the age at sexual maturity), see fig S19. Some of these regression models were further adjusted for average species weight (denoted lifespanAdjWeight).

For brevity, we will focus on linear regression models since phylogenetic regression models led to qualitatively similar conclusions (**tables S13–S14**). The most significant lifespan-related CpGs are located in the distal intergenic region neighboring *TLE4* (Pearson $r = 0.68$, $p = 5.8 \times 10^{-46}$, **Fig. 5A**, **table S11**) and two CpGs near the promoter region of *HOXA4* ($R = 0.66$, $p = 7.5 \times 10^{-45}$, midnight blue module, **Fig. 5A**), and negatively-correlated with a CpG in an intron of *GATA3* ($R = -0.65$, $p = 8.8 \times 10^{-42}$, **Fig. 5A**). Many of these significant CpGs remain so after phylogenetic adjustment, such as the CpGs neighboring *TLE4*, *HOXA4* ($p = 4.2 \times 10^{-5}$, $p = 4.8 \times 10^{-3}$ respectively, **fig. S17** and **table S11-S12**). The top 1,000 lifespan-related CpGs (comprising 500 positively and 500 negatively lifespan related CpGs) significantly overlapped (Fisher exact $p = 5.5 \times 10^{-134}$) with those found in our weight-adjusted analysis (lifespanAdjWeight).

In general, methylation of lifespan related CpGs does not change with age in mammalian tissues (**Fig. 5B**, **fig. S20**). The same can be seen from EWAS of lifespan restricted to animals of a given age group (e.g., only very young animals, **fig. S20D**). The EWAS of lifespan in all animals (irrespective of age) is highly correlated ($r > 0.7$) with the analogous EWAS restricted to animals that are young, middle-aged, or old, animals respectively.

EWAS of lifespan showed good consistency with the eigengene-based analysis in the mammalian co-methylation network. As expected, the previously discussed lifespan-related modules were enriched with CpGs implicated by our EWAS of lifespan: midnightblue (hypergeometric test $P = 2.2 \times 10^{-47}$; 67/249 overlapped CpGs), greenyellow (hypergeometric $P = 2.1 \times 10^{-36}$; 70/398 overlapped CpGs), tan (hypergeometric $P = 6.7 \times 10^{-23}$; 52/365 overlapped CpGs), and green (hypergeometric $P = 5.0 \times 10^{-18}$; 104/1542 overlapped CpGs) module.

In total, 1006 genes had a differential methylation association with lifespan (union of cognate genes resulting from the marginal model analysis for lifespan and lifespanAdjWeight). The gene expression levels of 17 of these genes exhibited a highly significant repeated measures correlation with chronological age (repeated measures Cor p value $< 10^{-50}$) in different human tissues (**Fig. 5C**). Two of these genes, *PTCHD4* and *ZBTB7B*, were also implicated by EWAS of weight-adjusted lifespan (lifespanAdjWeight). The cognate genes next to the top 500 positively lifespan-related CpGs play a critical role in animal organ morphogenesis (marginal model lifespan GREAT enrichment false discovery rate, $FDR = 3 \times 10^{-4}$ and LifespanAdjWeight $FDR = 3.3 \times 10^{-7}$, Fig. 5D), increased rib number in mice ($FDR = 1 \times 10^{-21}$, **Fig. 5D**), and implicates the HOXL gene group ($FDR = 0.004$ and weight adjusted LifespanAdjWeight $FDR = 1.3 \times 10^{-15}$), and abnormal survival in mice ($FDR < 4 \times 10^{-4}$).

Upstream regulators of maximum lifespan

We employed Ingenuity Pathway analysis {Kramer, 2014 #758} to identify potential upstream regulators of the genes cognate to the top 500 positively and top 500 negatively lifespan-related CpGs. The top-ranked candidate regulators of both gene lists included SOX2-OCT4-NANOG pluripotency factors ($FDR = 5.7 \times 10^{-4}$ lifespan negative, $FDR = 5.7 \times 10^{-4}$ lifespan positive), which play critical roles in cellular reprogramming. We performed a control analysis that ruled out potential confounding by sequence conservation (**fig. S25**). Upstream regulators also included several candidates related to development: sonic hedgehog (SHH), lifespan negative $FDR = 1.3 \times 10^{-4}$; POU4F2, lifespan negative $FDR = 3.3 \times 10^{-7}$ and ASCL1, lifespan negative $FDR = 1.6 \times 10^{-3}$ (**Fig. 5E**). These findings suggest that expression of lifespan-related genes might be regulated to some extent by pluripotency factors. This prompted us to investigate whether expression of any of the lifespan-related genes identified above are altered by transient expression of pluripotency inducing factors (Yamanaka factors OSKM) in a mouse model {Browder, 2022 #734}. Indeed, this analysis revealed that transient expression of OSKM altered the expression of 195 out of 646 lifespan-related genes in skin and 166 lifespan-related genes in the kidney (nominal Fisher exact $p = 3.9 \times 10^{-52}$ for skin and lifespan; $p = 1.4 \times 10^{-42}$ for kidney and lifespan, **Fig. 5F, fig. S32**). Genomic positions that are known to be bound by pluripotency factors (in at least one human/murine cell type according to ChIP-seq data from Encode) are located near CpGs that are associated with maximum species lifespans: NANOG binding sites are enriched for CpGs that are positively correlated with lifespan ($FDR = 0.002$) and to CpGs underlying the midnightblue module ($FDR = 0.0006$), which has high methylation levels in long-lived species (**Fig. 5G**). OCT4 (POU5F1) ($FDR = 0.02$) and cMYC ($FDR = 0.003$) binding sites are enriched with CpGs in the greenyellow module, which has low methylation levels in long-lived species (**Fig. 5G**). The ChIP-seq binding location analysis also implicates other noteworthy factors such as POLII, CTCF, RAD21, YY1, and TAF1, which show the strongest enrichment for negatively lifespan-related CpGs (**Fig. 5G**).

Given the role of CTCF in regulating the 3D organization of the genome, we conducted an enrichment analysis of topologically associating domain (TAD) boundaries and loop boundaries

identified in both human and mouse cell lines (**fig. S26**). We found that both TAD and loop boundaries demonstrated significant enrichment of negatively lifespan-related CpGs (FDR=3x10⁻⁴ for TAD boundaries and FDR=6.7x10⁻⁴ for loop boundaries in various cell lines, such as olfactory receptor cells, as well as human fibroblasts IMR90 and HFFc6; **fig. S26**).

CpGs linked to lifespan in various taxonomic orders and tissues

To pinpoint CpGs associated with log maximum lifespan independent of phylogenetic order or tissue type, we conducted a meta-analysis of EWAS findings from 25 distinct strata, comprising phylogenetic order and tissue type. Using a non-parametric meta-analysis approach (rankPvalue), we assessed the EWAS of lifespan (meta.lifespan) in these strata to identify CpGs unconfounded by tissue type or phylogenetic order (table S24). Our meta.lifespan results demonstrated significant overlap with the previously mentioned EWAS of lifespan in all eutherian species (hypergeometric $P = 1 \times 10^{-175}$, **Fig. 6A**). In contrast, none of the meta.lifespan CpGs overlapped with EWAS of age, which further support the statement that methylation of lifespan-related CpGs does not change with age in mammalian tissues. The top 4 CpGs from the meta.lifespan analysis are depicted in **Fig. 6B**, showing significant positive correlations for CpGs near LOXL1 and ZSCAN29 (exons), and negative correlations for those near RAB29 (exon) and GATA3 (downstream) with log maximum lifespan across various taxonomic orders and tissue types. Similar to our above mentioned results, CpGs implicated by our meta lifespan analysis (FDR<0.05) overlap significantly (FDR<0.01) with genes involved in organ morphogenesis, RNA biosynthesis, increased rib number in mice, Wnt signaling (**Fig. 6C**), and genes altered by transient expression of pluripotency-inducing factors in mouse models (nominal Fisher exact $p < 10^{-5}$ for skin and lifespan meta; $p < 10^{-11}$ for kidney and meta.lifespan, **Fig. 6D**).

Chromatin state analysis

Our large-scale mammalian DNAm data confirms that CpGs located in promoter regions (-2000 to 2000 bp of TSS regions) have low methylation levels (**Fig. 7A**, mean=15%). In contrast, those in gene bodies and distal regions are highly methylated (**Fig. 7A**, mean value ~70%). CpGs having a high/low mean methylation level tend to have positive/negative Z statistics for lifespan, respectively (**Fig. 7A-B**). We find that CpGs with low methylation levels in long-lived species are located close to the transcriptional start site of genes and near binding sites of Polycomb repressive complex 1 (PRC1, $p = 6.4 \times 10^{-11}$, **Fig. 7C**) and Polycomb repressive complex 2 (PRC2, $p = 2 \times 10^{-6}$). To test the hypothesis that long-lived species exhibit high/low methylation levels in chromosomal regions that are expected to have high/low methylation patterns, we used chromatin states that were identified and annotated based on over 1000 epigenetic data sets encompassing a diverse range of human cell and tissue types {Vu, 2022 #313}.

The lifespan related CpGs are enriched with transcriptional start site chromatin state (TSS1, $p = 2.5 \times 10^{-12}$), and flanking promoter states (PromF4, $p = 5.6 \times 10^{-10}$; PromF5, $p = 2.0 \times 10^{-9}$; PromF2, $p = 3.0 \times 10^{-4}$, **Fig. 7C**).

The CpGs with high methylation levels in blood samples of long-lived species are enriched in gene body associated states (notably transcribed state TxEx1, $p = 7.5 \times 10^{-8}$ and highly transcribed state TxEx4 $p = 1.7 \times 10^{-6}$, **Fig. 7D**). Detailed description of the chromatin state enrichment for EWAS of maximum lifespan is in the SupplementaryT.

A bi-clustering analysis between chromatin states and co-methylation modules based on fold enrichments (**Fig. 8; table S21; table S22**) revealed that the 55 mammalian co-methylation

modules fall into three large groupings (referred to as meta modules). The bar plot to the left of **Fig. 8** shows different mean methylation levels of the CpGs underlying the 3 meta modules: mean methylation=0.23, 0.66, and 0.77 for meta modules 1, 2, and 3, respectively.

Meta module 1 contains several chromatin states that are associated with Polycomb repression, including bivalent regulatory regions (BivProm1, 2) and ReprPC1. Further, meta module 1 contains chromatin states related to transcriptional start sites (TSS1, TSS2), and several flanking promoters (PromF2,3,4,5). TSS1, PromF2, and PromF4-5 (associated with negatively lifespan-related CpGs) were previously associated as the universal chromatin states with the strongest enrichments for CpG islands (54-101 fold) {Vu, 2022 #313}. The color band underneath **Fig. 8** reveals that six modules underlying meta-module 1 are sensitive to murine lifespan interventions. Meta module 1 is enriched with CpGs that have low methylation levels in long-lived species (overlap with EWAS of lifespan, tan/greenyellow modules, **Fig. 8**).

Meta-module 2 can be considered as a partially methylated module (mean methylation 0.66) and is enriched with several enhancer states, late replicating domains (partially methylated domains, common PMD {Zhou, 2018 #656}), and solo CpGs (WCGW, {Zhou, 2018 #656}). Meta-module 2 also contains the module most related with lifespan (midnightblue) and the human mortality risk module (magenta). These two modules overlap with the CpGs that are positively related to lifespan. Three out of four average weight-related modules are also located in meta-module 2.

Discussion:

In this study, we present an analysis of a cross-species DNA methylation dataset, obtained from a mammalian array platform. This platform specifically focuses on highly conserved regions of DNA, making it a valuable resource for studying methylation patterns across mammalian species {Arneson, 2022 #722}. The successful construction of mammalian phyloepigenetic trees suggests that the divergence of DNA methylation profiles is closely aligned with genetic changes throughout evolution. Sensitivity assessments reveal that the observed phyloepigenetic associations are not due to technical issues associated with our measurement platform. Instead, the phyloepigenetic signal may stem from sources like upstream regulators, transcription factors, or DNA sequence variations in distant regions.

The conserved CpGs exhibiting the strongest phylogenetic signals are situated in intergenic regions, while promoter regions do not display such signals. Previous studies report a rapid evolutionary rate of enhancers as a shared feature among mammalian genomes, while promoters demonstrate either full or partial conservation across species {Villar, 2015 #828}.

We found that 30 of the resulting 55 modules identified from an unsupervised machine learning method were readily associated with species traits (taxonomic order, maximum lifespan, average adult weight) or individual traits (chronological age, tissue, sex). We expect that many of the remaining 25 modules will be associated with biological characteristics about which we currently have no information. As a case in point, although the yellow module was not associated with any of our primary tested traits, it did show association with response to a murine circadian rhythm disruption study (light pollution during the night, **fig. S7B**). The upstream regulator analysis of the EWAS of lifespan identified the pluripotency transcription factors (OCT4, SOX2, and NANOG). We show that the transient overexpression of OSKM in murine tissues affects the methylation levels of CpGs near genes implicated by our EWAS of maximum lifespan (Fig. 5E). We speculate that the enhanced activity of the pluripotency network in long-lived species results in more efficient tissue repair and maintenance, ensuring a longer lifespan.

Both the EWAS and eigengene-based analyses identified methylation signatures of maximum lifespan presumably established at birth. Most of these were independent of aging and interventions that affect murine mortality risk. Several CpGs that are more highly methylated in long-lived species are located near *HOXL* genes and other genes that play a role in morphogenesis and development. Some of these lifespan related CpGs are located next to genes that are also implicated in our analysis of upstream regulators (e.g., *ASCL1* and *SMAD6*).

CpGs with methylation levels that are inversely related to lifespan are enriched in transcriptional start site (TSS1) and promoter flanking (PromF4, PromF5) associated chromatin states. Genes located in chromatin state TSS1 are constitutively active and enriched for nucleic acid metabolic processes {Vu, #313}. This could imply that long-lived species either evolved selective mechanisms to maintain low methylation levels near transcription start sites or may have adaptations that promote the high expression of essential genes. This high expression may indirectly prompt more active DNA demethylation mechanisms.

Method summary

The Mammalian Methylation Consortium generated cytosine methylation data from n=15,456 DNA samples derived from 70 tissue types of 348 mammalian species (331 eutherians, 15 marsupials, 2 monotremes) using a custom-designed mammalian methylation array that targets CpGs at conserved loci in mammals {Arneson, 2022 #722}. DNA methylation data were used for phyloepigenetic tree development using 1-cor dissimilarity applied to mean methylation values per species. The choice of the correlation-based dissimilarity matrix is justified in Supplementary Methods.

For unsupervised analysis, we formed WGCNA networks based on two sets of CpG probes in our data. The first network was generated from 14,705 conserved CpGs in 10,927 samples of 167 eutherian species. The preservation of this network was evaluated in an independent dataset comprising 3,692 samples from 29 tissues of 228 mammalian species (164 new species, 64 overlapped with the training set). The second network was a subset of 7,956 conserved CpGs in 11,105 samples from 167 eutherian and nine marsupial species. In addition, we developed seven consensus co-methylation networks to remove the confounding effects of species and tissue type. Consensus WGCNA can be interpreted as a meta-analysis across networks in different species and tissue types {Langfelder, 2008 #31; Langfelder, 2014 #275}.

For the eutherian network (Net1), module eigengenes (MEs) were defined as singular vectors (corresponding to the highest singular value) from the singular value decomposition of the scaled CpGs that underlie the respective module. The eigengenes in the eutherian network (Net1) explained a range of 24–63% (average = 43%) of the variance in the methylation data in the training set, replication set, and all data in each module (table S3). For a given module, we defined the measure of module membership, kME, as the Pearson correlation between the module eigengene and the CpGs. The association of module eigengenes was examined for different traits using individual regression models.

EWAS of lifespan was done in 28,318 CpGs that apply to mice and humans according to calibration/titration data (correlation with calibration exceeds 0.8) and mappability information as described in {Arneson, #722}. Since the distribution of maximum lifespan and other life history traits were highly skewed, we imposed a log-transformation on these phenotypes before conducting EWAS. Our tissue type specific EWAS was conducted in tissues with enough species (N>25 species) available. For our various EWAS of log transformed maximum lifespan, we

adopted a nominal significance threshold of 1.8×10^{-6} ($=0.01/28,318$) based on the conservative Bonferroni adjustment. We report a false discovery rate in our enrichment studies to adjust for multiple comparisons.

References and Notes

Acknowledgments:

Funding:

The majority of the funding was contributed by the Paul G. Allen Frontiers Group (Horvath) and Open Philanthropy (Horvath). Additional financial support for specific aspects of data generation was obtained from the following sources:

National Geographic Society grant 8941-11 (BL),

British Heart Foundation (FS/18/39/33684) (DV),

Wellcome (WT202878/Z/16/Z) (DTO),

European Research Council (788937) (DTO),

Cancer Research UK (20412) (DTO),

Max Planck Research Group Award from the Max Planck Gesellschaft (SCV)

UKRI Future Leaders Fellowship (MR/T021985/1) (SCV)

DST-NRF SARCHI chair of Mammalian Behavioural Ecology and Physiology (GUN 64756) (NCB)

Science Foundation Ireland Future Frontiers 19/FFP/6790 (ECT)

National Institute on Aging Intramural Research Program, NIH. (JAM), U01 AG060908 (Horvath), AG055841 (Mozhui), and AG043930/AG/NIA (Williams)

National Institute of Health AG065403 (Gladyshev), AG047200 (Gladyshev), AG067782 (Gladyshev) and AG076607 (Gladyshev), AG047200 (Gorbunova).

A.H. Schultz Foundation, Department of Evolutionary Anthropology, University of Zurich (Michael Krützen),

Leverhulme Trust RPG-2019-404 (Amy Pederson),

UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program (Jason Ernst)

Funds provided by the College of Computer, Mathematical and Natural Sciences at the University of Maryland, College Park (Gerald Wilkinson)

Taronga Conservation Society Australia (Jusine K. O'Brien)

Sonja C Vernes was supported by a Max Planck Research Group Award from the Max Planck Gesellschaft and a UKRI Future Leaders Fellowship (MR/T021985/1). Nigel Bennett was funded by a DST-NRF SARChI chair of Mammalian Behavioural Ecology and Physiology (GUN 64756). Collection of plains zebra samples was supported by the National Geographic Society to Brenda Larison (8941-11).

Author contributions:

Conceptualization: SH, AH

Formal Analysis: AH, CZL, ATL, QY, JZH

Investigation: All authors

Visualization: AH, CZL

Funding acquisition: SH, BL, DV, DTO, SCV, NCB, ECT, JAM

Project administration: SH

Supervision: SH

Writing – original draft: AH, CZL

Writing – review & editing: AH, SH, KR, JE and all other co-authors

Competing interests:

Steve Horvath, Adriana Arneson, and Jason Ernst are inventors on patent/patent application (Publication number WO2020150705) held/submitted by University of California Los Angeles that covers the mammalian methylation array technology.

S.H. and Robert Brooke are founders of the non-profit Epigenetic Clock Development Foundation, which has licensed several patents from UC Regents, and distributes the mammalian methylation array.

Data and materials availability:

All data from the Mammalian Methylation Consortium are posted on Gene Expression Omnibus (Complete dataset: GSE223748). Subsets of the data sets can also be downloaded from accession numbers, GSE174758, GSE184211, GSE184213, GSE184215, GSE184216, GSE184218, GSE184220, GSE184221, GSE184224, GSE190660, GSE190661, GSE190662, GSE190663, GSE190664, GSE174544, GSE190665, GSE174767, GSE184222, GSE184223, GSE174777, GSE174778, GSE173330, GSE164127, GSE147002, GSE147003, GSE147004). The mammalian array platform is distributed by the non-profit Epigenetic Clock Development Foundation (<https://clockfoundation.org/>).

The mammalian data can also be downloaded from the Clock Foundation webpage: <https://clockfoundation.org/MammalianMethylationConsortium>.

The manifest file of the mammalian array, genome annotations of the CpGs, and codes can be found on Github {Haghani, 2023 #926}.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S32

Tables S1 to S24

Data S1 to S19

References (45–95)

Fig. 1. Phyloepigenetic trees parallel the mammalian evolutionary tree. (A) The traditional phylogenetic tree from the TimeTree database {Kumar, 2017 #321} based on 321 (out of 348) species in our study. A full description of the species in our study is reported in **table S1**. (B) Blood-based phyloepigenetic tree created from hierarchical clustering of DNA methylation data in this study (additional analysis in fig. S3A,B). We formed the mean value per cytosine across samples for each species. The clustering used 1 minus the Pearson correlation (1-cor) as a pairwise dissimilarity measure and the average linkage method as intergroup dissimilarity. Phyloepigenetic trees for skin and liver can be found in fig. S2. Additional analyses, e.g., involving different choices of CpGs or intergroup dissimilarity measures, are reported in the supplement (**fig. S2**). The colored bars reflect the branch height. (C) Scatter plot of the distances in blood phyloepigenetic (1-cor) vs the traditional evolutionary tree. (D) Scatter plots displaying the log-odds ratios of regions exhibiting phylogenetic signals relative to the transcription start site (TSS) are presented. The phylogenetic signal is determined using Blomberg's K statistic {Blomberg, 2003 #826}. In this analysis, CpGs were grouped into categories using sliding windows relative to the TSS, ensuring a minimum count of 10 CpGs per group. To assess enrichment, the Fisher exact overlap test was employed, focusing on the top 500 CpGs displaying phylogenetic signals within each region. The results indicate notable enrichment ($OR > 3$) in certain intergenic and genic regions, but not in promoters. Additional analysis in **fig. S4**.

Fig. 2. DNA methylation network relates to species and individual characteristics in mammalian species. (A) the WGCNA network of 14,705 conserved CpGs in eutherian species (Network 1). The identified modules related to species, or individual sample characteristics. Network 1 modules were compared to eight additional networks (fig. S5). The modules with strong associations with species and sample characteristics were labeled below the dendrogram. Grey color codes CpGs that are outside of modules. (B) summary of the modules that showed strong associations with species and individual sample characteristics. The +/- labels are the direction of association with each trait. (C) Top defined functional biological processes related to network 1 modules (details in **fig. S9, table S4**). (D) mammalian co-methylation modules form clusters of proteins in the STRING protein-protein interaction (PPI) network. For the sake of visualization, the analysis was limited to the top 50 CpGs with the highest module membership value per module. colors: mammalian network 1. The lollipop plot shows the global cluster coefficient {Barrat, 2004 #819} of the proteins within a module (up to 500 top CpGs) in a PPI network. Our permutation analysis matched the distribution of the original module sizes. We evaluated 1100 random permutations, i.e. 20 for each of the 55 modules. The boxplot reports the global clustering coefficient per module (y-axis) versus permutation status: module resulting from a random selection of proteins (left) versus original module resulting from WGCNA (right). The modules with cluster coefficients larger than the maximum permutation cluster coefficient were considered

as significant at $p=0.001$. The dashed vertical line corresponds to the maximum global clustering coefficient observed in the 1100 random permutations.

Fig. 3. Co-methylation modules related to mammalian maximum lifespan, weight, human mortality, and age. Modules associated with log maximum lifespan ($p<10^{-20}$) (A) or log average species weight ($p<10^{-17}$) (B) in marginal association: correlation test with the mean module eigengene of the species. The module eigengene is defined as the 1st principal component of the scaled CpGs underlying a module. The species are randomly labeled by their animal number (table S1). (C) The top modules associated with median life expectancy, upper limit life expectancy, or average adult weight of 93 dog breeds, model: marginal correlation test of the mean module eigengene with target variables (detailed breed characteristics are in table S8). R, Pearson correlation coefficient; p, correlation test p-value. (D) Forest plots of the top modules associated with mortality risk in the Framingham Heart Study Offspring Cohort (FHS), and Women's Health Initiative (WHI) study, totaling 4651 individuals (1095, 24% death). The right panel indicates the number of deaths/total number of individuals in each study. We report the meta-analysis p-value in the title of the forest plot. (E) Module that correlates significantly ($p<1\times 10^{-300}$) with relative age (defined as ratio of age/maximum lifespan) across mammalian species using a multivariate regression model. Covariates: tissue, sex, and species differences. Each dot corresponds to a eutherian tissue sample ($n=14,542$). Dots are colored by taxonomic order. (F) The volcano plot of the rmCorrelation of all purple module genes in GTEx data (Additional analysis in fig. S11).

Fig. 4. The effects of different pro-aging and anti-aging interventions on selected DNAm modules. Six DNA methylation modules respond to lifespan-related intervention experiments and are associated with the life expectancy of the mouse models. In contrast, the mammalian maximum lifespan modules do not correspond directly to the benefits or stress triggered by the intervention in the murine samples. (A) Changes in the intervention modules in the liver parallel smaller size and longer life expectancy of growth hormone receptor mouse models (GHRKO). Sample size: GHRKO, 11 (5 female, 6 male); Wt, 18 (9 male, 9 female). The age range: 6–8 months. (B) Caloric restriction (CR) DNA methylation module signature predicts longer lifespan in this treated group. Age=18 months; Sex=Male; N=CR, 59; control, 36. (C) High-fat diet accelerates aging in the age module. N=high-fat diet, 133 (125 females, 8 males); control (*ad libitum*), 212 (202 females, 10 males). Age range: 3–32 months. (D), (E) Examining the effects of *in vivo* partial reprogramming on intervention modules. (D) a schematic view of the partial programming experiment in 4F mice {Browder, 2022 #734}. A systemic Yamanaka factors expression (Oct4, Sox2, Klf4, Myc) was periodically induced by adding doxycycline to drinking water for two days per week. The partial programming was done at three different durations. Sample size: control (C57BL/6+dox), $n=7$; 1m 4F, $n=3$; 7m 4F, $n=5$; 10m 4F, $n=3$ (all tissues except skin, $n=2$ for skin). (E) scatter plots of the linear changes of the intervention modules in the skin and kidney of mice treated with different durations (dosages) of Yamanaka factors. R, Pearson correlation coefficient; p, correlation test p-value. The intervention modules indicate a dose-dependent rejuvenation of skin and kidney by this partial programming regimen.

Fig. 5. Epigenome-wide association study (EWAS) of mammalian log-transformed maximum lifespan. (A) The figure represents the CpG-specific association with maximum lifespan across $n=333$ eutherian species. For EWAS, the mean methylation values of each CpG (per species) were regressed on log maximum lifespan. The right portion of the panel reports EWAS results after adjustment for average adult weight. Genome annotation: human hg19. Red

dotted line: Bonferroni corrected two-sided p-value $< 1.8 \times 10^{-6}$. The point colors indicate the corresponding modules. The bar plot indicates the top enriched (hypergeometric test, eutherian probes as background) modules for the top 1000 (500 negative CpGs nominal $p < 1.1 \times 10^{-11}$, FDR = 1×10^{-10} ; 500 negative CpGs positive CpGs nominal $p < 1.5 \times 10^{-21}$, FDR = 7.5×10^{-20}) significant CpGs for different EWAS. **(B)** Venn diagram of the overlaps between top hits from EWAS of maximum lifespan and meta-analysis of age (meta-analysis results from {Lu, 2021 #744}, additional analysis in fig. S20). **(C)** Venn diagram of the overlaps between the genes adjacent to the EWAS results and top age-related mRNA changes in human tissues ($p < 1 \times 10^{-50}$). **(D)** Gene set enrichment analysis of the genes proximal to CpGs associated with mammalian maximum lifespan. We only report enrichment terms that are significant after adjustment for multiple comparisons (hypergeometric FDR < 0.01) and contain at least five significant genes. The top three significant terms per column (EWAS) and enrichment database are shown in the panel. **(E)** Ingenuity potential upstream regulator analysis {Kramer, #758} of the differentially methylated genes related to mammalian maximum lifespan. **(F)** Venn diagram of 3 gene lists. First, the top 646 genes adjacent to 1000 lifespan related CpGs (500 positive and 500 negative). Gene lists 2 and 3 are based on CpGs that are differentially methylated (nominal Wald test $p < 0.005$, up to 500 positive and 500 negatively related CpGs) after OSKM overexpression in murine kidney (601 genes) and skin (695 genes) {Browder, 2022 #734}. We observe significant overlap between the gene lists (nominal Fisher exact $p = 3.9 \times 10^{-52}$ for skin and lifespan; $p = 1.4 \times 10^{-42}$ for kidney and lifespan) **(G)** transcriptional factor motif enrichment analysis of lifespan modules and lifespan related CpGs. The enrichment results for LifespanAdjWeight.negative were not significant. The overlap is assessed by a hypergeometric test for the CpGs within the motifs based on the human hg19 genome.

Fig. 6. CpGs Linked to Lifespan in Various Taxonomic Orders and Tissues. Using the non-parametric rankPvalue method {Langfelder, 2008 #31}, we combined 25 EWAS of lifespan results from various taxonomic order or tissue type strata, calculating the significance of a CpG's consistently high (or low) rank based on the 25 EWAS of log maximum lifespan (meta lifespan, underlying EWAS results can be found in table S24, Data S19). **(A)** The overlap of top 1000 (500 per direction) meta-lifespan CpGs with EWAS of lifespan in all eutherians (nominal Fisher exact $p = 1 \times 10^{-175}$). **(B)** Scatter plots illustrating the top meta-lifespan CpGs categorized into different tissue-phylogenetic order strata are presented. Each panel displays only the strata that exhibit significant relationships. The first row represents the phylogenetic order strata combining all tissues. **(C)** Gene set enrichment analysis of the genes proximal to CpGs associated with mammalian maximum lifespan. We only report enrichment terms that are significant after adjustment for multiple comparisons (hypergeometric FDR < 0.01) and contain at least five significant genes. The top three significant terms per column (EWAS) and enrichment database are shown in the panel. **(D)** Venn diagram of 3 gene lists. First (the bottom circle), the top 407 genes adjacent to 1000 meta-lifespan CpGs (500 positive and 500 negative). Gene lists 2 and 3 (the top circles) are based on CpGs that are differentially methylated (nominal Wald test $p < 0.005$, up to 500 positive and 500 negatively related CpGs) after OSKM overexpression in murine kidney (601 genes) and skin (695 genes) {Browder, 2022 #734}.

Fig. 7. Chromatin state analysis and distance to the transcriptional start site for the lifespan related CpGs. **(A)** The illustrated plot present mean methylation across species (displayed on left y-axis) and EWAS of maximum lifespan Z-statistics (shown on right y-axis), all plotted against

the distances to the closest transcription start site (represented on the x-axis). **(B)** Mean methylation across species (y-axis) is plotted against EWAS Z statistics for log maximum lifespan in different genomic regions (intergenic, promoter, gene body). Additional EWAS results after adjustment for phylogenetic relationships can be found in **fig. S17-20** and corresponding enrichment results can be found in **fig. S22-S24**. Pearson correlation coefficients and p-values are reported in different panels. Chromatin state enrichment analysis of **(C)** the top 500 negatively lifespan related CpGs, and **(D)** top 500 positively lifespan related CpGs. The columns in each panel correspond to EWAS results for log transformed maximum lifespan across i) all tissues combined (Lifespan.All), ii) blood samples only (Lifespan.Blood), iii) skin samples only (Lifespan.Skin), meta-lifespan and the corresponding results after adjustment for average adult weight, Lifespan (AdjWeight). The last column reports enrichment with respect to the relativeAge⁽⁺⁾ module (purple). We use the same significance thresholds as in Figure 5. The cells' shadings correspond to fold enrichment between co-methylation modules and each chromatin state. The cells' numeric values correspond to the p-value of such enrichments based on the hypergeometric test, and only cells' values with significant p-value<0.001 (equivalent to FDR<0.02) are shown. The chromatin states are learned based on epigenetic datasets profiling chromatin mark signals in different human cell and tissue types, resulting in a genome annotation shared across cell types {Vu, 2022 #313}. The common partially methylated domains (commonPMD), solo CpGs (WCGW), and highly methylated domain (HMD) annotations are from {Zhou, 2018 #656}. Polycomb repressor complexes (PRC) 1 and PRC 2 binding sites are obtained from the ChIP-seq datasets of PRC 1 and 2 from ENCODE (ENCODE Project Consortium, 2012){Consortium, #823}.

Fig. 8. Mammalian methylation meta modules based on the chromatin states and external genome annotations. The heatmap shows the enrichments between (1) mammalian co-methylation modules and significant lifespan related EWAS CpG groups (x-axis) EWAS and (2) chromatin states or other genomic annotation (y-axis). The cells' shadings correspond to log transformed fold-enrichment values (observed count divided by expected count). Hypergeometric tests were used to evaluate the enrichment significance in each cell, and * indicates a nominal p-value<0.001 (FDR<0.10). Only chromatin states and external genome annotations with at least one significant enrichment (FDR<0.10) are shown. The chromatin states are based on a human based universal chromatin annotation of human cell and tissue types {Vu, 2022 #313}. Other genomic annotations include the common partially methylated domains (commonPMD), solo CpGs (WCGW), and highly methylated domain (HMD) annotations, which are from {Zhou, 2018 #656}. In addition, Polycomb repressor complexes (PRC) 1 and PRC 2 binding sites are defined from the ChIP-seq data of PRC 1 and 2 from ENCODE (ENCODE Project Consortium, 2012){Consortium, #823}. The row and column hierarchical clustering trees (average linkage) are based on a dissimilarity (1 minus the pairwise Pearson correlation between log transformed fold enrichment values). The left side barplot indicates the mean methylation levels of the CpGs in each state for all eutherian samples in our data. We used the 14,705 eutherian CpGs as the background for enrichment of the co-methylation modules. In contrast, 28,318 CpGs (high quality probes in humans and mice) were used as a background for enrichment of significant lifespan related EWAS CpG groups with chromatin states and genome annotations. Each EWAS CpG

group includes up to 500 most significant CpGs per direction (positively/negatively related with lifespan) as detailed in the caption of **Fig. 5**.