UNIVERSITY OF CALIFORNIA SAN DIEGO

Characterizing the Evolution of Epigenetic Clocks at Different Time Scales

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Tina Wang

Committee in charge:

Professor Trey Ideker, Chair
Professor Peter Ernst
Professor Bing Ren
Professor Elizabeth Winzeler
Professor Kun Zhang

2019

The Dissertation of Tina Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

This dissertation is dedicated to my parents, Rong-Qi Wang and Shu Guang Xu, for their everlasting support of my professional endeavors. To Brandon Santos, my loving husband who has been the best thing that's ever happened to my life. To Belli(ni), my dog, without whom, this work would have never occurred.

TABLE OF CONTENTS

LIST OF SUPPLEMENTARY FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my mentor, Trey Ideker, for giving me the opportunity to learn under his guidance. He supported my vision to pursue projects that were outside the scope of his lab's interests. With his mentorship, I became an expert in computational biology and improved my communication skills.

I would also like to thank Anne-Ruxandra Carvunis. Anne was my personal mentor during my first two years of graduate school and a role model. She showed immense care and guidance that were essential towards my development as an independent researcher. Her passion inspired my own interests in evolutionary biology.

I would also like to thank the Genetics Training Program for providing funding, and Bruce Hamilton for the insightful journal clubs with amazing bagels. I would also like to thank my committee: Peter Ernst, Bing Ren, Elizabeth Winzeler, and Kun Zhang, for their feedback on my dissertation. I would especially like to thank Elizabeth Winzeler for her candid conversations and everlasting support for women in science.

I would also like to thank members of Trey Ideker's group. Andy Gross who inspired the last two chapters of my dissertation. Michael Yu who gave me guidance in both technical and personal matters. Justin Huang who helped me achieve fantasy football success, and more importantly, kept me grounded throughout our journey together. Jason Kreisberg who taught me to purge oxford commas and the use of Arial from my work. Jianzhu Ma for his contribution towards the development of my third chapter. Jisoo Park for her refreshing views towards science and enjoyment of Korean BBQ. Charlotte Curtis who was instrumental towards navigating Trey's schedule. Katherine Licon for being the real-life Chuck Norris. Barry Demchak for his patient

guidance in all things cluster and great email threads. Erica Silva, Brenton Munson and Samson Fong, for their encouragement and companionship on the lab-side.

I would like to thank my parents, Rong-Qi Wang and Shu Guang Xu, for providing the resources and encouragement to pursue my interests. They taught me the meaning of teamwork, persistence and determination. Xiao-Yu Wang for always leading the way to my professional development. My husband, Brandon Santos, who made sure that I could overcome my challenges by providing companionship, nourishment and organization to my life. My dog, Belli(ni), for being my scientific muse.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Evidence for a common evolutionary rate in metazoan transcriptional networks" in *eLife*, 2015 by Anne-Ruxandra Carvunis, Tina Wang, Dylan Skola, Alice Yu, Jonathan Chen, Jason F Kreisberg and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Evidence for a common evolutionary rate in metazoan transcriptional networks" in *Genome Biology*, 2017 by Tina Wang, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission of the material as it may appear as “A conserved epigenetic progression aligns dog and human age” by Tina Wang, Jianzhu Ma, Andrew Hogan, Samson Fong, Brian Tsui, Jason F. Kreisberg, Peter D. Adams, Anne-Ruxandra Carvunis, Danika Bannasch, Elaine A. Ostrander and Trey Ideker. The dissertation author was the primary investigator and author of this material.

VITA

2011	University of California Berkeley *Bachelor of Science, Molecular Toxicology*
2019	University of California San Diego *Doctor of Philosophy, Biomedical Sciences*

PUBLICATIONS

Adam E. Field, Neil A. Robertson, **Tina Wang**, Aaron Havas, Trey Ideker, and Peter D. Adams. “DNA Methylation Clocks in Aging: Categories, Causes, and Consequences.” Molecular Cell 71, no. 6 (September 20, 2018): 882–95. https://doi.org/10.1016/j.molcel.2018.08.008.

Nam Bui, Justin K. Huang, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, **Tina Wang**, Wei Zhang, John Paul Shen, Jason F. Kreisberg and Trey Ideker “Disruption of NSD1 in Head and Neck Cancer Promotes Favorable Chemotherapeutic Responses Linked to Hypomethylation.” Molecular Cancer Therapeutics 17, no. 7 (July 2018): 1585–94. https://doi.org/10.1158/1535-7163.MCT-17-0937.

Tina Wang, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Michael Ku Yu, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams and Trey Ideker. “Epigenetic Aging Signatures in Mice Livers Are Slowed by Dwarfism, Calorie Restriction and Rapamycin Treatment.” Genome Biology 18, no. 1 (March 28, 2017): 57. https://doi.org/10.1186/s13059-017-1186-2.

John J. Cole, Neil A. Robertson, Mohammed Iqbal Rather, John P. Thomson, Tony McBryan, Duncan Sproul, **Tina Wang**, Claire Brock, William Clark, Trey Ideker, Richard R. Meehan, Richard A. Miller, Holly M. Brown-Borg and Peter D. Adams. “Diverse Interventions That Extend Mouse Lifespan Suppress Shared Age-Associated Epigenetic Changes at Critical Gene Regulatory Regions.” Genome Biology 18, no. 1 (March 28, 2017): 58. https://doi.org/10.1186/s13059-017-1185-3.

Anne-Ruxandra Carvunis*, **Tina Wang***, Dylan Skola*, Alice Yu, Jonathan Chen, Jason F. Kreisberg and Trey Ideker. “Evidence for a Common Evolutionary Rate in Metazoan Transcriptional Networks.” eLife 4 (December 18, 2015). https://doi.org/10.7554/eLife.11615.

Erika L. Flannery, **Tina Wang**, Ali Akbari, Victoria C. Corey, Felicia Gunawan, A. Taylor Bright, Matthew Abraham, Juan F. Sanchez, Meddly L. Santolalla, G. Christian Baldeviano, Kimberly A. Edgel, Luis A. Rosales, Andrés G. Lescano, Vineet Bafna, Joseph M. Vinetz and Elizabeth A. Winzeler. “Next-Generation Sequencing of Plasmodium Vivax Patient Samples Shows Evidence of Direct Evolution in Drug-Resistance Genes.” ACS Infectious Diseases 1, no. 8 (August 14, 2015): 367–79. https://doi.org/10.1021/acsinfecdis.5b00049.

Loes M. Olde Loohuis, Jacob A. S. Vorstman, Anil P. Ori, Kim A. Staats, **Tina Wang**, Alexander L. Richards, Ganna Leonenko, James T. Walters, Joseph DeYoung, GROUP consortium, Rita M. Cantor and Roel A. Ophoff. "Genome-Wide Burden of Deleterious Coding Variants Increased in Schizophrenia." Nature Communications 6 (July 9, 2015): 7501. https://doi.org/10.1038/ncomms8501.

S. A. J. de With, S. L. Pulit, **T. Wang**, W. G. Staal, W. W. van Solinge, P. I. W. de Bakker and R. A. Ophoff. "Genome-Wide Association Study of Lymphoblast Cell Viability after Clozapine Exposure." American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 168, no. 2 (2015): 116–22. https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.b.32287.

Nongluk Plongthongkum, Kristel R. van Eijk, Simone de Jong, **Tina Wang**, Jae Hoon Sul, Marco P. M. Boks, René S. Kahn, Ho-Lim Fung, Roel A. Ophoff and Kun Zhang. "Characterization of Genome-Methylome Interactions in 22 Nuclear Pedigrees." PloS One 9, no. 7 (July 14, 2014): e99313. https://doi.org/10.1371/journal.pone.0099313.

Andrew Taylor Bright, Micah J. Manary, Ryan Tewhey, Eliana M. Arango, **Tina Wang**, Nicholas J. Schork, Stephanie K. Yanow and Elizabeth A. Winzeler. "A High Resolution Case Study of a Patient with Recurrent Plasmodium Vivax Infections Shows That Relapses Were Caused by Meiotic Siblings." PLoS Neglected Tropical Diseases 8, no. 6 (June 2014): e2882. https://doi.org/10.1371/journal.pntd.0002882.

Justin M. Richner, Karen Clyde, Andrea C. Pezda, Benson Yee Hin Cheng, **Tina Wang**, G. Renuka Kumar, Sergio Covarrubias, Laurent Coscoy and Britt Glaunsinger. "Global mRNA Degradation during Lytic Gammaherpesvirus Infection Contributes to Establishment of Viral Latency." PLoS Pathogens 7, no. 7 (July 2011): e1002150. https://doi.org/10.1371/journal.ppat.1002150.

Bifeng Yuan, Jing Zhang, Hongxia Wang, Lei Xiong, Qian Cai, **Tina Wang**, Steven Jacobsen, Sriharsa Pradhan and Yinsheng Wang. "6-Thioguanine Reactivates Epigenetically Silenced Genes in Acute Lymphoblastic Leukemia Cells by Facilitating Proteasome-Mediated Degradation of DNMT1." Cancer Research 71, no. 5 (March 1, 2011): 1904–11. https://doi.org/10.1158/0008-5472.CAN-10-3430.

*co-first authors

ABSTRACT OF THE DISSERTATION

Characterizing the Evolution of Epigenetic Clocks at Different Time Scales

by

Tina Wang

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2019

Professor Trey Ideker, Chair

A fundamental concept of molecular biology is that cellular functions are shared among all living organisms. Identifying DNA sequences that are conserved across species are generally thought as a proxy towards uncovering important cellular functions. Such proxies have not been appropriate to understand the functional consequences of epigenetic modifications, which regulate the expression of protein-coding genes. Unlike the DNA sequence, epigenetic modifications are dynamic and change depending on factors including cellular context and age. To further elucidate the importance of these epigenetic modifications towards functional effects and relationship with DNA sequence, I characterized epigenetic changes while taking into account the differing dynamics with respect to time.

In the first chapter, I asked if different life histories that yield differences in genomic architecture are also reflected in the rate of epigenomic evolutionary changes by comparing the rate of epigenemic evolution in closely related species of *Drosophila* and *Mus musculus*. I found that, despite large differences in genomic architecture, the rate of epigenetic evolution was strikingly similar, such that they reflect the molecular clock that was observed in protein evolution.

For the remainder of my thesis, I studied the changes of a particular epigenetic modification, DNA methylation, with respect to a single lifetime across species. Previous studies in humans have demonstrated that DNA methylation can be used to measure age in humans, reflecting an epigenetic clock. In the second chapter, I asked whether these changes can also measure age in mice. I found that epigenetic clocks can also measure age in mice, and that these clocks can also be slowed in long-lived mice. In my third chapter, I asked whether these changes with respect to age are conserved among mammals. For this purpose, I specifically characterized methylation changes with respect to age in conserved sequences between dogs and humans. I found that these changes were conserved and can be used to translate age across mammals.

Overall, my work characterizes the evolution of epigenetic modifications across species, which seemingly act as clocks measuring both evolutionary time and chronological age.

INTRODUCTION

All living organisms contain cells, the fundamental unit of all biological studies. The genome, DNA, of the cell encodes the information that enables cellular function via transcription of RNA and translation into protein. Biological function is believed to be created by stochastic processes, through naturally occurring mutations that are selected through evolution. If a mutation is beneficial for an individual, then this mutation will propagate with each generation of progeny, and eventually becomes fixed within a population. In contrast, detrimental mutations would be expected to be removed from the population due to the forces of selection. Moreover, selection varies for distinct lineages. For instance, smaller species are subjected to higher levels or predation. As such, these species tend to have shorter lifespan, reprodroduce faster and have more progeny than larger species. These differences also can be seen in the genomes of species as detrimental variants are removed faster in short-lived species relative to long-lived species during evolution. For these reasons, if a sequence is observed in many distinct lineages, then it likely encodes an important function for cellular fitness.

This conceptual framework helped our understanding of the protein-coding sequences in the genome, which are generally conserved from yeast to mammals. However, these sequences only represent 2% of the human genome. Moreover, these sequences are nearly identical between human and chimpanzees (99%), indicating that regulation of protein-coding genes is important for the differences observed (King and Wilson 1975). Distinguishing between nonfunctional and functional sequences remains challenging. If we simply evaluate sequences based on their presence across diverse mammals, we would find that that up to 10% of the human genome is under selective constraint, meaning only 8% can be regarded as ‘functional’ (Siepel et al. 2005). This approach may not be entirely appropriate for determining whether sequences are functional for regulation of protein-coding genes. Proteins are produced from the sequence itself, but regulation of protein-coding genes is accomplished by biochemical, or epigenetic, modifications of DNA. The epigenome is comprised of several distinct modifications including histones, transcription factors and DNA methylation that determine the expression of genes. Recent studies profiling chromatin states have found evidence that up to half of the human genome is covered in epigenetic modifications (ENCODE Project Consortium 2012). Yet, the relationship between the regulation of protein-coding genes and these modifications are still unknown. Epigenetic modifications are frequently found outside of highly-conserved sequences. Therefore, requiring sequence constraint may not be enough to determine the function of these modifications.

The epigenome is dynamic, changing throughout life, and contrasts to the static genomic sequence that encodes these changes. The changes in the epigenome due to cellular environments allow the formation of different cell types from the same genomic sequence. During embryogenesis, the epigenome is completely reprogrammed enabling the formation of a multicellular organism from a single cell, the zygote (Gifford et al. 2013). Unlike the sequence that encodes a functional protein, the sequences that enable this epigenetic regulation are likely functional in a context-dependent manner (Kellis et al. 2014). For these reasons, how we define function must be with respects to the requirements of a dynamic system.

We can still apply comparative evolutionary approaches to improve our understanding of the epigenome. Since the static genomic sequences likely determine the functions of the epigenome, we can compare across species and evaluate similarities with respect to these dynamic processes. In this work, I utilize a comparative evolutionary approach to improve our understanding of these epigenetic modifications by characterizing these changes across species at different scales of time, over evolutionary time and over a single lifespan. When we consider evolutionary scales of time, we can identify functionally active regions in closely related species (Villar et al. 2015; Stefflova et al. 2013; He et al. 2011; Bradley et al. 2010; Paris et al. 2013) or delineating broad trends underlying regulatory networks when comparing distantly related species (Boyle et al. 2014; Won et al. 2013; Cheng et al. 2014; Schmidt et al. 2010). In my first chapter, I leverage naturally occurring differences of genomic evolution in mammals and fruit flies to determine whether these differences also correspond to differences in epigenome evolution. This strategy will further elucidate whether sequence constraints are also echoed in the epigenomic constraints. If the dynamics differ, this could suggest that there could be other constraints acting on the evolution of the epigenome. In chapters 2 and 3, I utilize comparisons across species at their respective lifespans to understand the extent that mammalian genomes similarly experience aging regardless of the large differences in lifespan. In chapter 2, I ask whether a specific epigenetic modification, DNA methylation, can measure age in mice, as it was previously described and validated in human populations (Hannum et al. 2013; Horvath 2013; Gross et al. 2016). After confirming this epigenetic correlate of age across mammals, I then ask whether these changes are conserved during aging in dogs. Overall, this body of work uses comparative epigenomic approaches, identifying similar changes that reflect molecular epigenetic clocks, and change across species in similar manners when examining over evolutionary time or over a span of a lifetime.

References

Boyle, Alan P., Carlos L. Araya, Cathleen Brdlik, Philip Cayting, Chao Cheng, Yong Cheng, Kathryn Gardner, Ladeana W. Hillier, Judith Janette, Lixia Jiang, Dionna Kasper, Trupti Kawli, Pouya Kheradpour, Anshul Kundaje, Jingyi Jessica Li, Lijia Ma, Wei Niu, E. Jay Rehm, Joel Rozowsky, Matthew Slattery, Rebecca Spokony, Robert Terrell, Dionne Vafeados, Daifeng Wang, Peter Weisdepp, Yi-Chieh Wu, Dan Xie, Koon-Kiu Yan, Elise A. Feingold, Peter J. Good, Michael J. Pazin, Haiyan Huang, Peter J. Bickel, Steven E. Brenner, Valerie Reinke, Robert H. Waterston, Mark Gerstein, Kevin P. White, Manolis Kellis, and Michael Snyder. 2014. “Comparative Analysis of Regulatory Information and Circuits across

Distant Species." *Nature* 512 (7515): 453–56.

Bradley, Robert K., Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A. Tonkin, Mark D. Biggin, and Michael B. Eisen. 2010. "Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species." *PLoS Biology* 8 (3): e1000343.

Cheng, Yong, Zhihai Ma, Bong-Hyun Kim, Weisheng Wu, Philip Cayting, Alan P. Boyle, Vasavi Sundaram, Xiaoyun Xing, Nergiz Dogan, Jingjing Li, Ghia Euskirchen, Shin Lin, Yiing Lin, Axel Visel, Trupti Kawli, Xinqiong Yang, Dorrelyn Patacsil, Cheryl A. Keller, Belinda Giardine, Mouse ENCODE Consortium, Anshul Kundaje, Ting Wang, Len A. Pennacchio, Zhiping Weng, Ross C. Hardison, and Michael P. Snyder. 2014. "Principles of Regulatory Information Conservation between Mouse and Human." *Nature* 515 (7527): 371–75.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.

Gifford, Casey A., Michael J. Ziller, Hongcang Gu, Cole Trapnell, Julie Donaghey, Alexander Tsankov, Alex K. Shalek, David R. Kelley, Alexander A. Shishkin, Robbyn Issner, Xiaolan Zhang, Michael Coyne, Jennifer L. Fostel, Laurie Holmes, Jim Meldrim, Mitchell Guttman, Charles Epstein, Hongkun Park, Oliver Kohlbacher, John Rinn, Andreas Gnirke, Eric S. Lander, Bradley E. Bernstein, and Alexander Meissner. 2013. "Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells." *Cell* 153 (5): 1149–63.

Gross, A. M., P. A. Jaeger, J. F. Kreisberg, K. Licon, K. L. Jepsen, M. Khosroheidari, B. M. Morsey, S. Swindells, H. Shen, C. T. Ng, K. Flagg, D. Chen, K. Zhang, H. S. Fox, and T. Ideker. 2016. "Methylome-Wide Analysis of Chronic HIV Infection Reveals Five-Year Increase in Biological Age and Epigenetic Targeting of HLA." *Molecular Cell* 62 (2): 157–68.

Hannum, G., J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J. B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Molecular Cell* 49 (2): 359–67.

He, Qiye, Anaïs F. Bardet, Brianne Patton, Jennifer Purvis, Jeff Johnston, Ariel Paulson, Madelaine Gogol, Alexander Stark, and Julia Zeitlinger. 2011. "High Conservation of Transcription Factor Binding and Evidence for Combinatorial Regulation across Six Drosophila Species." *Nature Genetics* 43 (5): 414–20.

Horvath, S. 2013. "DNA Methylation Age of Human Tissues and Cell Types." *Genome Biology* 14 (10): R115.

Kellis, Manolis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, Ewan Birney, Gregory E. Crawford, Job Dekker, Ian Dunham, Laura L. Elnitski, Peggy J. Farnham, Elise A. Feingold, Mark Gerstein, Morgan C. Giddings, David M. Gilbert, Thomas R. Gingeras, Eric D. Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D. Lieb, Richard M. Myers, Michael J. Pazin, Bing Ren, John A.

Stamatoyannopoulos, Zhiping Weng, Kevin P. White, and Ross C. Hardison. 2014. “Defining Functional DNA Elements in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6131–38.

King, M. C., and A. C. Wilson. 1975. “Evolution at Two Levels in Humans and Chimpanzees.” *Science* 188 (4184): 107–16.

Paris, Mathilde, Tommy Kaplan, Xiao Yong Li, Jacqueline E. Villalta, Susan E. Lott, and Michael B. Eisen. 2013. “Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression.” *PLoS Genetics* 9 (9): e1003748.

Schmidt, Dominic, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P. Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T. Odom. 2010. “Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding.” *Science* 328 (5981): 1036–40.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. 2005. “Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes.” *Genome Research* 15 (8): 1034–50.

Stefflova, Klara, David Thybert, Michael D. Wilson, Ian Streeter, Jelena Aleksic, Panagiota Karagianni, Alvis Brazma, David J. Adams, Iannis Talianidis, John C. Marioni, Paul Flicek, and Duncan T. Odom. 2013. “Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals.” *Cell* 154 (3): 530–40.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, Robert Deaville, Jonathan T. Erichsen, Anna J. Jasinska, James M. A. Turner, Mads F. Bertelsen, Elizabeth P. Murchison, Paul Flicek, and Duncan T. Odom. 2015. “Enhancer Evolution across 20 Mammalian Species.” *Cell* 160 (3): 554–66.

Won, Kyoung-Jae, Xian Zhang, Tao Wang, Bo Ding, Debasish Raha, Michael Snyder, Bing Ren, and Wei Wang. 2013. “Comparative Annotation of Functional Regions in the Human Genome Using Epigenomic Data.” *Nucleic Acids Research* 41 (8): 4423–32.

CHAPTER 1: Evidence for a common evolutionary rate in metazoan transcriptional networks

1.1 Abstract

Genome sequences diverge more rapidly in mammals than in other animal lineages such as birds or insects. However, the effect of this rapid divergence on transcriptional network evolution remains unclear. Recent reports have indicated a faster divergence of transcription factor binding in mammals than in insects, but others found the reverse for mRNA expression levels. Here, we show that these conflicting interpretations resulted from differing methodologies between lineages. We performed an integrated analysis of transcriptional network evolution by examining mRNA expression, transcription factor binding and *cis*-regulatory motifs across >25 animal species including mammals, birds and insects. Strikingly, we found that transcriptional networks evolve at a common rate over time across the three animal lineages. Furthermore, differences in rates of genome divergence were greatly reduced when restricting comparisons to chromatin-accessible sequences. The evolution of transcription is thus decoupled from the global rate of genome sequence evolution, suggesting that a small fraction of the genome regulates transcription.

1.2 Introduction

A long-standing question in biology is what fraction of the genome regulates transcription (Consortium 2012, Graur et al. 2013, Niu and Jiang 2013, Kellis et al. 2014). Recent studies of chromatin structure have implicated half of the human genome in regulatory interactions (Consortium 2012). Comparative genomic studies, however, have shown that less than 10% of the human genome is evolutionarily conserved (Siepel et al. 2005), suggesting that many of the experimentally-detected interactions are not functional (Graur et al. 2013). Recent studies have measured the association between sequence changes and changes in transcript levels, epigenetic modifications or binding of transcription factors regulating specific gene sets (gene-specific transcription factors, GSTF) (Cookson et al. 2009, McVicker et al. 2013, Kasowski et al. 2013, Kasowski et al. 2010, Heinz et al. 2013, Villar, Flicek, and Odom 2014, Wong et al. 2015, Brem et al. 2002, Shibata et al. 2012, Chan et al. 2009). These experiments demonstrated that genomic sequences can influence transcription even in the absence of evolutionary conservation. For instance, some repetitive elements previously thought to be "junk" DNA have been shown to effectively regulate gene expression (Rebollo, Romanish, and Mager 2012). The rapid evolution of repetitive and other rapidly-evolving sequences could cause pervasive rewiring of transcriptional networks through creation and destruction of regulatory motifs (Villar, Flicek, and Odom 2014). Such rapid transcriptional evolution would set mammals apart from other metazoans like birds or insects, whose genomes contain far fewer repetitive elements (Taft, Pheasant, and Mattick 2007) and tend to be more constrained (Siepel et al. 2005, Zhang et al. 2014).

A few studies have attempted to assess whether transcriptional networks evolve more rapidly in mammals than in insects from the fruit fly genus *Drosophila*. These studies have reached conflicting conclusions. When examining the evolution of GSTF binding, chromatin immune-precipitation (Wiens et al.) studies in mammalian livers have generally described faster divergence rates than similar studies in fly embryos (Villar, Flicek, and Odom 2014, Stefflova et al. 2013). However, divergence rates were estimated with different analytical methods in the different ChIP studies (**Supplementary Table S1.1**) (Bardet et al. 2012, Villar, Flicek, and Odom 2014). Another study found that gene expression levels may diverge at a slower rate in mammals than in flies, by comparing genome-wide correlations of mRNA abundances estimated by RNA sequencing (RNA-seq) for mammals but by a mixture of technologies for flies including microarrays (Coolon et al. 2014). Although the inconsistencies between these conclusions may indicate that the evolution of transcriptional networks is fundamentally different in mammals and insects, they may also reflect a sensitivity of evolutionary rate estimations to technical methodology.

Here, we jointly examined the evolution of gene expression levels and the underlying genome-wide changes in GSTF binding and *cis*-regulatory sequences using consistent methodologies both within and across various animal lineages.

1.3 Results

We assembled a comparative genomics platform encompassing >40 publicly available datasets spanning >25 organisms representative of the *Mammalia* (mammals), *Aves* (birds) and *Insecta* (insects) phylogenetic classes (**Supplementary Figure S1.1**). We designed a statistical framework, applicable to virtually any type of genomic data, to objectively compare the rates of divergence of various data types across lineages. In brief, an exponential model describing evolutionary divergence under a common, lineage-naïve rate was evaluated against a lineage-aware model, accounting for both statistical significance and effect size (**Figure 1.1**). We assessed the power of this statistical framework using simulations and found that it could detect differences in divergence rates with high sensitivity (**Methods**; **Supplementary Figure S1.2**).

As a baseline, we first performed a comparative analysis of the evolution of genome sequences. We randomly sampled genomic segments from designated reference genomes: *Mus musculus domesticus* (C57BL/6) for mammals, *Gallus gallus* for birds and *Drosophila melanogaster* for insects. The rates at which genomic segments that homology with the other species within each lineage accumulate nucleotide substitutions were then estimated and compared using our statistical framework. Segments retaining homologs displayed high sequence conservation across all three lineages, although our framework detected a slightly but significantly faster divergence in insects than in mammals or birds (P < 0.05; **Supplementary Figure S1.3**). Next, we compared the rates at which randomly sampled genomic segments lost homology with the other species within each lineage. We observed a much larger difference in evolutionary rates across lineages using this measure ($P < 0.05$; **Figure 1.2**; **Supplementary Figure S1.4**). For instance, after 100 million years (Myrs) of evolution, only ~30% of mammalian segments retained homology whereas >60% of bird and insect segments did. These findings recapitulated previous observations according to which genome sequences are less constrained in mammals than in insects (Siepel et al. 2005) or birds (Zhang et al. 2014)

We then studied the evolution of gene expression levels, using exclusively RNA-seq datasets. In mammals and birds these datasets were generated from adult livers; in insects, they were from whole bodies of adult female fruit flies (**Methods; Supplementary Table S1.2**). After determining expression levels for each gene in each species using a common data processing pipeline, we correlated the expression levels of genes in the reference species with the expression levels of their one-to-one orthologs in all other species within the same lineage (Methods). We found that correlations of gene expression levels decreased over time at similar rates that were statistically indistinguishable: a lineage-naïve model describing the evolution of gene expression levels under a common rate fitted the data as well as a lineage-aware model (**Figure 1.3**). This result was robust to changes in correlation metrics or inclusion/exclusion of poorly expressed genes (**Supplementary Figure S1.5**).

Several lines of evidence suggest that gene expression levels can remain relatively stable even as the genomic locations bound by GSTFs change rapidly over time (Paris et al. 2013, Wong et al. 2015, Chan et al. 2009). Therefore, we next examined the evolution of GSTF binding patterns. We considered all GSTFs that were profiled using ChIP followed by massively parallel sequencing (ChIP-seq) in at least three related species, where separate ChIPs were performed per species. GSTFs meeting these requirements were Twist and Giant in fruit fly embryos, and CEBPA, FOXA1 and HNF4A in mammalian livers (**Methods**; **Supplementary Table S1.1; Supplementary File S1.1**). We aimed to measure cross-species similarity in GSTF occupancy with a unified analytical method across all of these datasets. Despite the widespread use of ChIP-seq, there is no consensus on the appropriate analytical method (Wilbanks and Facciotti 2010). ChIP-seq analysis pipelines typically discretize continuous occupancy profiles into a set of occupied segments ("peaks"), but this step requires choosing a signal processing algorithm (a peak caller) and associated parameters (**Figure 1.4a**). Further comparison of occupied segments across species requires additional analytical choices (**Figure 1.4a**), some of which can strongly influence downstream findings (Bardet et al. 2012).

To explore the impact of these choices, we processed all ChIP-seq data using systematic combinations of parameters representative of, and expanding from, previous studies (**Supplementary Table S1.1**) (Landt et al. 2012). In total, we executed 108 analytical pipelines to compare divergence rates across 6 pairs of GSTFs (2 in insects each compared with 3 in mammals), the occupancy profiles of which were examined in 3 – 7 species per lineage (Methods). The values of the estimated rates varied greatly from one combination of parameters to the next (**Figure 1.4b, c**). However, in the majority of cases (56 – 78% over the 6 comparisons), GSTF binding patterns diverged at statistically indistinguishable rates in mammals and insects (**Figure 1.4d; Supplementary File S1.2**). Although the computed divergence rates were sensitive to technical methodology (**Supplementary Figure S1.6**), for a given method the results were generally similar across lineages for all of the five GSTFs investigated.

To substantiate these findings, we devised a method to compare genome-wide occupancy profiles at single-nucleotide resolution without discretization. We correlated occupancy profiles between pairs of species across all nucleotides where genomes aligned, after accounting for the differences in sequencing depth, read length and fragment size across datasets (Methods). Again, we found indistinguishable divergence rates, regardless of which GSTF or lineage was examined (**Figure 1.4e**). After 100 Myrs of evolution, the correlation of GSTF occupancy profiles was 0.10 in mammals and 0.13 in insects. As a control, we also applied this method to CTCF, a pleiotropic DNA-binding protein that acts as chromatin insulator and looping factor (Ohlsson, Lobanenkov, and Klenova 2010). In mammals, patterns of DNA occupancy have been shown to be more conserved for CTCF than for GSTFs using unified analytical methods (Schmidt et al. 2012). In contrast, CTCF DNA occupancy was shown to diverge rapidly in insects, perhaps due to the existence of other insulator proteins (Ni et al. 2012, Villar, Flicek, and Odom 2014). Our analysis successfully recapitulated this difference (**Figure 1.4f**), demonstrating that the common evolutionary rate observed among GSTFs (**Figure 1.4e**) was not an artifact of our method for profile correlation.

The similarity of divergence rates observed across lineages for gene expression levels (**Figure 1.3**) and GSTF binding patterns (**Figure 1.4**) was unexpected given the rapid evolution of genomic sequences in mammals relative to insects (Siepel et al. 2005) or birds (Zhang et al. 2014) (**Figure 1.2**). We therefore further examined these trends at the level of *cis*-regulatory sequences. First, we considered the DNA sequence motifs thought to be specifically recognized by the mammalian and insect GSTFs included in the previous ChIP-seq analysis (**Figure 1.4**). We identified locations with significant matches to these motifs throughout the genomes of the reference species and estimated how frequently these loci retained the same motifs relative to background expectations (**Methods**). We found similar, indistinguishable retention rates in mammals and insects (**Figure 1.5a**). Next, we studied the evolution of a broader set of motifs corresponding to GSTFs shared between *M. musculus* and *D. melanogaster*. We found that these motifs were retained at similar rates across lineages relative to background expectations in 8 out of 12 cases (one example shown in **Figure 1.5b**; all other cases in **Supplementary Figure S1.7**).

Most active *cis*-regulatory sequences are located in genomic regions with accessible chromatin (Hesselberth et al. 2009). A recent study showed that chromatin-accessible sequences were significantly more conserved between human and mouse than expected by chance (Yue et al. 2014). We expanded this analysis to a wide range of species by using chromatin-accessible sequences identified by DNAse I hypersensitivity in *M. musculus* livers, *D. melanogaster* embryos and *G. gallus* MSB-1 cells **(Methods)**. We performed the segment sampling procedure described previously (**Figure 1.2**), after excluding genes and promoter regions since they typically are highly conserved **(Methods)**. Whereas inaccessible segments lost homology much faster in mammals than in insects and birds ($P < 0.05$; **Figure 1.5c**), accessible segments retained homologs at more similar rates in the three lineages (**Figure 1.5d; Supplementary Figure S1.8**). We still detected statistically significant differences across lineages ($P < 0.05$), but the effect sizes were considerably smaller than for inaccessible segments. For instance, ~60% of segments retained homology after 100 Myrs in birds and insects, independently of accessibility, whereas ~50% of chromatin-accessible segments and only ~20% of inaccessible segments did so in mammals.

1.4 Discussion

To our knowledge, the analyses presented here represent the most comprehensive study conducted to date on the evolution of transcriptional networks across animal lineages. By applying unified analytical methods to data from different lineages, we were also able to glean novel insights into the evolution of transcriptions in animals. We observed that gene expression levels, GSTF binding patterns, regulatory motifs and chromatin-accessible sequences each diverged at rates that were similar across mammals, birds and insects. These unexpected results reconcile previously conflicting findings (Coolon et al. 2014, Villar, Flicek, and Odom 2014), highlighting the importance of unified study methodologies and providing evidence for a common evolutionary rate in metazoan transcriptional networks.

Most functional genomics studies have focused on humans and model organisms such as *D. melanogaster* or *M. musculus*, which are distantly related to each other. However, data on closely related species, like those which we collected in this study, are needed to investigate the dynamics of molecular network evolution. Unfortunately, such data remain scarce, leading to important limitations of our work. We only investigated three lineages and six to twelve organisms per lineage with non-uniform coverage over evolutionary time. In addition, we only examined a small number of tissues for each lineage and a total of five GSTFs (none in birds). The generalizability of our observations thus remains to be further evaluated as more data becomes available. Despite these limitations, our finding that transcriptional networks evolve at a common rate per year across animal lineages was strikingly robust across data layers.

The underlying mechanisms responsible for this concordance of evolutionary rates are unclear. Mammals, birds and insects exhibit wide differences in the features that are traditionally associated with evolutionary rates, such as generation times and breeding sizes. Populations with small breeding sizes, such as mammals, are thought to be more prone to genetic drift (Ohta 1992). This theory accounts for the abundance of repetitive elements and the rapid evolution of genomic sequences in mammals relative to insects, which have much larger breeding sizes. If the same theoretical principles also governed the evolution of transcriptional networks, we would have expected that transcription would evolve more rapidly in mammals than in insects. Instead, our results show that the evolution of transcriptional networks, whether slow (e.g. transcript levels) or fast (e.g. GSTF binding), is decoupled from the lineage-specific features that govern genome sequence evolution.

One potential model could be that repetitive and rapidly-evolving sequences, which make up the majority of the mammalian genome (Siepel et al. 2005, Taft, Pheasant, and Mattick 2007), play a negligible role in the global regulation of gene expression. Rather, chromatin-accessible regions may represent the only portion of the mammalian genome that effectively regulates transcription. We observed that chromatin-accessible regions diverge much more slowly than other non-coding sequences in mammals, consistent with previous findings (Yue et al. 2014). These differences in divergence rates, however, were not found in birds and insects. As a result, chromatin-accessible regions in mammals are conserved at levels similar to those in birds and insects, in contrast to the genome as a whole. According to this model, the similar rates of evolution of chromatin-accessible sequences would constrain the dynamics of transcriptional evolution to be similar across lineages. The regulatory potential of repetitive and other rapidly-evolving elements could be rendered functionally inconsequential by silencing, or could be concentrated on controlling the expression of genetic elements that we did not investigate, such as non-coding RNAs or species-specific genes (Sundaram et al. 2014).

An alternative model could be that the sequences that control transcriptional regulation in birds and insects evolve particularly rapidly within otherwise stable genomes. In these organisms, transcriptional networks would diverge under the action of natural selection, through specific single nucleotide substitutions resulting in rapid compensatory turnover (He, Holloway, et al. 2011). In mammals, transcriptional networks would diverge in a largely neutral fashion driven for instance by transposable elements (Sundaram et al. 2014). In this case, similar rates of transcriptional divergence across lineages would arise through very different evolutionary processes.

Importantly, none of the aforementioned models account for the differences in generation times between lineages. Evolutionary changes occurring based on chronological time and not generation time has also been observed for many protein-coding sequences. Observations such as these led to the molecular clock theory (Kumar 2005). The mechanisms through which environmental forces entrain these chronological evolutionary clocks remain to be elucidated (Kumar 2005).

1.5 Methods

Genome and Annotation Sources. We downloaded genome sequences for organisms belonging to three metazoan lineages: mammals, birds and insects. The mammalian and insect genome sequences were downloaded from the UCSC Genome Bioinformatics website (Rosenbloom et al. 2015): mm9 for *Mus musculus domesticus*, rn5 for *Rattus norvegicus* and hg19 for *Homo sapiens*; dm3 for *Drosophila melanogaster*, droSim1 for *Drosophila simulans*, droEre2 for *Drosophila erecta*, droYak2 for *Drosophila yakuba*, droAna3 for *Drosophila ananassae* and dp4 for *Drosophila pseudoobscura*. Genomes for mice strains and species not available from the

UCSC Genome Bioinformatics site (*M. musculus domesticus* (AJ), *M. musculus castaneus* and *M. spretus*) were downloaded from (Stefflova et al. 2013). We downloaded bird genome sequences from Ensembl version 80 BioMart (Cunningham et al. 2015): galGal4 for *Gallus gallus*, Turkey_2.01 for *Meleagris gallopavo*, taeGut3.2.4 for *Taeniopygia guttata* and FicAlb_1.4 for *Ficedula albicollis*. Protein-coding gene names and symbols along with associated transcripts sequences were obtained from FlyBase (dos Santos et al. 2015) for insect species (dmel-r5.46, dsim-r1.4, dere-r1.3, dyak-r1.3, dana-r1.3 and dpse-r2.30), from Ensembl version 80 BioMart for bird species and from Ensembl version 59 BioMart for mammalian species (Cunningham et al. 2015). For *M. spretus* and *M. musculus castaneus*, we used the same transcript annotations as for *M. musculus*. Within the genomes of our designated reference organisms (*M. musculus domesticus*, *G. gallus* and *D. melanogaster*), we defined promoters as 2kb upstream of transcription start site and delineated intergenic regions as regions that did not overlap annotated genes or promoters. Chromatin accessibility tracks used in **Figure 1.5c-d** and **Supplementary Figure S1.8** were downloaded from the UCSC bioinformatics website (Rosenbloom et al. 2015) for *M. musculus domesticus* and *D. melanogaster* and obtained from (He et al. 2014) for *G. gallus*. We restricted our analyses to the sequences or annotations in, or homologous to, the well defined chromosome scaffolds of the reference organism. Specific reference chromosomes analyzed are as follows: *G. gallus* (1-28, Z, W), *D. melanogaster* (2L, 2R, 3L, 3R, 4, X) and *M. musculus* (1-19, X, Y).

Homology and Evolutionary Relationships. We obtained orthology relationships between protein-coding genes using Ensembl COMPARA (Vilella et al. 2009), matching the Ensembl versions used for protein coding genes for each species described above. These relationships were used in **Figure 1.3** and **Supplementary Figure S1.5**. Homology between genomic segments was assigned using the LiftOver tool (Rosenbloom et al. 2015), for all analyses presented in **Figures**

1.2, 1.4 and 1.5 and associated figure supplements, with the exception of the nucleotide-resolution analysis of GSTF occupancy profiles presented in **Figure 1.4e-f**. We used pre-computed chain files from UCSC matching the genome versions listed above when chains were readily available (Rosenbloom et al. 2015). When chain files were not available, we built chain files to map the UCSC M. *musculus* C57BL/6 mm9 to the genomes of *M. musculus domesticus* AJ, *Mus musculus castaneus* and *Mus spretus*, as well as to map the Ensembl 80 galGal4 to the genomes of *M. gallopavo*, *F. albicollis* and *T. guttata* (**Supplementary Table S1.2). ** These chains were constructed by flowing the steps recommended by UCSC (**Supplementary Table S1.3**) (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto).

For the nucleotide-resolution analysis of GSTF occupancy profiles, we assigned homology relationships using the chain files, or, in the case of mice strains, using genome mapping tables from (Stefflova et al. 2013). We filtered the chain files to obtain one-to-one unambiguous mappings by retaining only highest scoring alignment for each position. These filtered mappings were then used to transfer data to from any organism onto the corresponding reference genome. Regions in the reference species genome lacking one-to-one unambiguous mappings were excluded from analysis.

To define evolutionary distances separating species in Myrs, we chose published estimates generated as homogeneously as possible within each lineage using a combination of sequence alignments and fossil records. All distances between insect species were taken from (Tamura, Subramanian, and Kumar 2004); all distances between bird species were taken from (Lu et al. 2015); distances between mammalian species were taken from (Stefflova et al. 2013) and TimeTree (Hedges 2009).

Data Sources. For RNA-seq analyses (**Figure 1.3**; **Supplementary Figure S1.5**), sequencing data for the reference species corresponding to two experiments performed independently by different research groups, and, when possible, representing different genotypes, were downloaded from public repositories. For *M. musculus domesticus*, we used data from (Goncalves et al. 2012, Sugathan and Waxman 2013), for *G. gallus* we used the data from (Brawand et al. 2011) and (Coble et al. 2014); for *D. melanogaster* we pulled from (Chen et al. 2014, Consortium 2012). Other species included were *M. musculus castaneus* (Goncalves et al. 2012) *M. spretus* (Wong et al. 2015), *R. norvegicus* (Gong et al. 2014), *H. sapiens* (Lin et al. 2014, Consortium 2012), *G. gorilla (Brawand et al. 2011)*, *D. simulans* (Chen et al. 2014), *D. yakuba* (Chen et al. 2014), *D. ananassae* (Chen et al. 2014), *D. pseudoobscura* (Chen et al. 2014), *M. gallopavo* (Monson et al. 2014), *A. platyrhynchos* (Huang et al. 2013), and *F. albicollis* (Uebbing et al. 2013). Specific accession numbers are listed in **Supplementary Table S1.2**.

For ChIP-seq analyses (**Figure 1.4**), we downloaded data for FOXA1 in *M. musculus domesticus* (C57BL/6) (Stefflova et al. 2013), *M. musculus domesticus* (AJ) (Stefflova et al. 2013), *M. musculus castaneus* (Stefflova et al. 2013), *M. spretus* (Stefflova et al. 2013) and *R. norvegicus* (Stefflova et al. 2013); HNF4A and CEBPA in *M. musculus domesticus* (C57BL/6) (Stefflova et al. 2013), *M. musculus domesticus* (AJ) (Stefflova et al. 2013), *M. musculus castaneus* (Stefflova et al. 2013), *M. spretus* (Stefflova et al. 2013), *R. norvegicus* (Stefflova et al. 2013), *H. sapiens (Schmidt et al. 2010)* and *C. familiaris (Schmidt et al. 2010)*; Twist in *D. melanogaster (He, Bardet, et al. 2011)*, *D. simulans (He, Bardet, et al. 2011)*, *D. erecta (He, Bardet, et al. 2011)*, *D. yakuba (He, Bardet, et al. 2011)*, *D. ananassae (He, Bardet, et al. 2011)* and *D. pseudoobscura (He, Bardet, et al. 2011)*; Giant *in D. melanogaster* (Bradley et al. 2010, Paris et al. 2013), *D. yakuba* (Bradley et al. 2010) and *D. pseudoobscura (Paris et al. 2013)*. We also gathered data for CTCF in *M. musculus domesticus* (C57/B6) (Schmidt et al. 2012), *R. norvegicus* (Schmidt et al. 2012), *H. sapiens* (Schmidt et al. 2012), *C. familiaris* (Schmidt et al. 2012), *D. melanogaster (Ni et al. 2012)*, *D. simulans (Ni et al. 2012)*, *D. yakuba (Ni et al. 2012)* and *D. pseudoobscura (Ni et al. 2012)*. Accession numbers corresponding to the specific experimental replicates and control samples are listed in **Supplementary File S1.1**.

For motif analyses (**Figure 5a-b**; **Supplementary Figure S1.7**), we gathered known position-weight matrixes from the JASPAR database (Mathelier et al. 2014) and the Fly Factor survey (Zhu et al. 2011). We focused on the motifs corresponding to Twist and Giant in *D. melanogaster*, to CEBPA, HNF4A and FOXA1 in *M. musculus domesticus*, and on a set of 12 other motifs corresponding to GSTFs conserved across mammals and insects. This set was constructed by downloading all Core A vertebrata motifs from JASPAR (Mathelier et al. 2014), identifying those corresponding to conserved GSTFs with one-to-one orthologs between *M. musculus domesticus* and *D. melanogaster* using COMPARA (Vilella et al. 2009), and filtering the list down to those 12 instances where a position-weight matrix was also described in Fly Factor (Zhu et al. 2011) and were not already analyzed.

Comparing evolutionary rates. We developed a statistical framework to compare evolutionary rates of various 'omics data layers between lineages, and implemented it in R (R Development Core Team 2011). This framework takes as inputs: measures of pairwise cross-species similarity (*e.g,* correlation of gene expression or sequence conservation), pairwise cross-species evolutionary distances and lineage labels..Conceptually, the framework estimates both a statistical significance and an effect size to determine whether rates of evolutionary divergence are indistinguishable or different between lineages (**Figure 1.1**).

In practice, we model evolutionary divergence by an exponential decay in log-linear space. First, the nls function is applied to the log-transformed cross-species similarity data as a function of evolutionary distances to derive the following linear models:

- a lineage-naïve model that estimates a shared intercept and slope for all the data without specifying the lineage labels
- a lineage-aware model that estimates a shared intercept for all the data and lineage-specific slopes based on lineage labels
- lineage-specific models that estimate intercept and slope individually for each lineage

Second, an R function written in-house to handle nls model structures estimates the significance level of an ANOVA with a likelihood ratio test comparing the lineage-naïve and the lineage-aware model. Third, we define the effect size as the predicted absolute difference in similarity between lineage pairs after 100 Myrs of divergence as estimated from the lineage-specific models. We consider that the framework detected a difference between evolutionary divergence rates when the significance level is <0.05 and the effect size is >0.05.

We chose to use an exponential decay function because it is the simplest evolutionary model that fit all our input measures of cross-species similarity reasonably well. We chose to model the exponential decay in log-linear space because we noted that a simple exponential decay in linear space failed to capture the conservation observed between distant species (mouse versus human at 91 Myrs and dog at 97.4 Myrs) when analyzing the evolutionary dynamics of GSTF binding (**Figure 1.4**) and motif retention (**Figure 1.5**) in mammals. We hypothesize that these data layers likely follow a more complex decay model, but we did not want to explore this with our current data set to avoid over-fitting.

The power of this statistical framework was assessed by simulating data for two lineages with measure of cross-species similarity decaying exponentially at different rates over time (**Supplementary Figure S1.2**). We fixed one lineage to decay at set rates: -0.007, -0.005 and -0.003. We fixed the second lineage to be faster by a range of given differences**.** Over 1,000 simulations, we sampled two values from a normal distribution centered on the expected values from the set exponential decay rates corresponding to the evolutionary distances shown in **Figure 1.4b,** with standard deviations set at 0.5% or 5%**.** Our framework detected an absolute rate difference of 0.001 at 39.3% of simulations and an absolute rate difference of 0.003 in 88.9% of simulations when the standard deviation was high (5%). When the standard deviation was low (0.5%), our framework detected an absolute rate difference of 0.001 in 25.7% of simulations and an absolute rate difference of 0.003 in 100% of simulations.

Gene expression evolutionary rates (related to Figure 1.4)**.** Analysis of gene expression evolutionary rates was performed in four steps. First, we preprocessed the raw RNA sequencing data downloaded for public data sources. Second, we quantified the abundance of all annotated transcripts corresponding to protein-coding genes. Third, we estimated cross-species similarity by correlating transcript abundances at the genome-scale. Finally, we used these cross-species similarity estimates as input to our statistical framework to evaluate a common model against a lineage-aware model.

RNA sequencing data was first preprocessed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic (Bolger, Lohse, and Usadel 2014). In order to quantify transcript abundances, we then used the program Sailfish (Patro, Mount, and Kingsford 2014) to 1) build transcriptome indices for each species using the transcriptome sequences described above, using the parameters "-p 8 -k 20"; 2) quantify transcript abundance using the transcriptome indices with the parameters “-p 8 -l ‘T=PE:O=><:S=U’ “ for samples with paired-end reads and “-p 8 –l ‘T=SE:S=U’ ” for samples with single-end reads. The bias-corrected transcripts per million (TPM) abundances estimated by Sailfish were then summed over the transcripts corresponding to the same gene locus.

To estimate cross-species similarities in gene expression levels, for each lineage, we used R (R Development Core Team 2011) to build a matrix containing the gene expression values for all the protein-coding genes of the reference organism and their one-to-one orthologs across other organisms within each lineage. We discarded instances where the abundance of a particular gene locus was less than or equal to 5 TPM. We then calculated the Spearman’s rank correlation for the expression of all genes between the reference and all other organisms within each lineage and plotted these correlations as against the evolutionary distance separating each organism pair (**Figure 1.3**). We also repeated the calculations using Kendall’s rank correlation coefficient and Pearson’s product-moment correlation on log$_2$-transformed expression values (**Supplementary Figure S1.5a, b**). Finally, we calculated Spearman’s correlations among all genes including those with less than 5 TPM (**Supplementary Figure S1.5c**). All these scenarios were evaluated using our statistical framework. None indicated that a lineage-aware model described the data better than a common model.

GSTF Occupancy – Segment-resolution (related to Figure 1.4a-d). The first step of all our occupancy analyses was to align the ChIP-seq reads to the corresponding genomes in order to obtain occupancy profiles (**Figure 1.4a**). For each accession (**Supplementary Table S1.1**), the sequencing reads were aligned to reference genomes using Bowtie2 version 2.2.4 (Langmead and Salzberg 2012) with the parameters “-very-sensitive -N 1.” Reads containing the 'XS:' field (multi-mappers) were removed. Reads having the same start site were presumed to be PCR duplicates and removed using the "rmdup" command of SAMtools version 1.1 (Li et al. 2009). The filtered reads were then converted to tagAlign format. The tagAlign files corresponding to CEBPA, HNF4A, FOXA1, Twist and Giant were then processed using 108 different segment-resolution methods and one nucleotide-resolution method; the tagAlign files corresponding to CTCF were only processed using the nucleotide-resolution method. The nucleotide-resolution method is described below and relates to **Figure 1.4e-f**.

The aim of our segment-resolution analyses was to examine how robust the evolution of GSTF binding patterns was across 108 different analysis pipelines (**Figure 1.4a-d**). We implemented all these pipelines, which follow the same general framework and differ only in the choice of 5 parameters, described and underlined below.

First, the occupancy profiles in the tagAlign files were discretized into candidate occupied segments using a peak caller algorithm that aims at identifying segments where the ChIP sample is enriched in reads relative to the control sample. We implemented two peak callers: MACS version 2 (M) (Zhang et al. 2008) and SPP (S) (Kharchenko, Tolstorukov, and Park 2008).

The occupied segments were then selected from the candidate set using a quality filter: stringent (S), lenient (L) or asymmetric (A). When using MACS2 (Zhang et al. 2008) as a peak caller, lenient segments were called using a p-value cutoff of 10^{-5} (default) and merged across replicates when available using the merge function in BEDTools (Quinlan and Hall 2010). Stringent segments were called using a p-value cutoff of 10^{-22} and intersected across replicates when replicates were available. The intersection procedure, inspired from (Stefflova et al. 2013), used BEDTools (Quinlan and Hall 2010) to implement the following two steps: 1) merge the two replicates 2) select the merged segments corresponding to at least one segment in each original replicate. When using SPP (Kharchenko, Tolstorukov, and Park 2008) as a peak caller, lenient segments were called using a q-value of 10^{-2} (default), and merged across replicates when available (Quinlan and Hall 2010). Stringent segments were called by selecting all candidate segments assigned to the lowest possible q-value in the sample, then intersected across replicates when available using the same intersection procedure. The asymmetric quality filter, inspired by (He, Bardet, et al. 2011, Bardet et al. 2012), indicates that segments were called stringently in the reference species and leniently in the other organism.

The coordinates of the occupied segments called in the reference organism were projected onto the other organism's genome using the LiftOver tool from the UCSC genome browser (Rosenbloom et al. 2015) and specifying a sequence similarity filter through the minMatch parameter. We used 3 different minMatch thresholds: stringent (S: 0.95 default), lenient (L: 0.5), and none (N: 0.001).

After cross-species coordinate projection, a reference subset was chosen to define the set of reference-occupied segments that would be further analyzed. Three choices were implemented: all reference-occupied segments independently of whether they map to any other species (A); for each pair of species, only reference-occupied segments with a homolog in the second species (P); only reference-occupied segments that had homologs across all the other species considered within the lineage (S).

The projected coordinates of the reference subset were then overlapped with the coordinates of the occupied segments in the other species using the intersect function in BEDTools (Quinlan and Hall 2010). The overlap requirement was either lenient (L; default parameter of 1 bp) or stringent (S; required a reciprocal overlap of half of the segments length: " -f 0.5 -r").

We systematically executed all combinations of the aforementioned 2 peak callers, 3 quality filters, 3 sequence similarity filters, 3 reference subsets, and 2 overlap requirements, yielding a total of 108 pipelines. The output of each pipeline was the fraction of reference subset segments that overlapped segments occupied in the others species (*i.e.* segments retaining occupancy between the two species). This output was used as a cross-species similarity measure for GSTF binding patterns. We analyzed these similarity measures for 6 pairs of GSTFs (Twist and Giant were each compared to FOXA1, CEBPA and HNF4A) using our statistical framework. Two GSTFs were considered to diverge differently from each other over time when 1) the significance of the test was less than 0.05 and 2) the effect size was greater than 0.05. In summary we found that the choice of parameters greatly influenced what the evolutionary dynamics of a given GSTF looked like (**Figure 1.4b-c**) but that in general the rate of divergence of mammal and insects GSTFs were statistically indistinguishable (**Figure 1.4d**). The results of these tests for all GSTF pairs considered across 108 pipelines are reported in **Supplementary File S1.2** and summarized as pie-charts in **Figure 1.4**. Observations about general trends of parameters and evolutionary divergence are further elaborated in **Supplementary File S1.3**.

GSTF Occupancy – Nucleotide-resolution (related to Figure 1.4e-f). In order to compare occupancy profiles directly without discretizing them into occupied segments and unoccupied segments, we correlated sets of imputed fragment density vectors across species. The inputs to this method were the tagAlign files described above. To generate these vectors we first estimated the mean fragment size using a method adapted from (Kharchenko, Tolstorukov, and Park 2008), whereby the mean fragment size is computed as the number of base pairs of offset between the positive and negative strands that maximizes the Pearson correlation coefficient of their mapped read density. We used a modified approach that considered only the density of 5' read start sites on each strand, rather than the density of the entire read. The first peak of the cross-correlation values was identified by approximating the first derivative by the finite difference method, smoothing the derivative values with a Gaussian kernel of bandwidth 10, and identifying the first downward zero-crossing of the curve. This position was used as the estimated mean fragment size *L*. We created imputed fragments by extending each read start site by *L* base pairs in the 3' direction. We then calculated a fragment density vector for each chromosome as the number of such imputed fragments that overlap each genomic position. When multiple replicates were available, replicates were merged by adding the fragment density vectors.

In order to minimize bias introduced by the presence of unmappable regions, we implemented a masking scheme that adaptively normalizes each dataset depending on the read length and estimated fragment size of each sequencing run. First, all possible error-free reads of a given length were generated synthetically and aligned back to the genome using Bowtie2 2.2.4 with the following parameters: "-r -N 0 -D 0 -R 0 --dpad 0 --score-min `C,0,-1`". Any multi-mapping reads with the 'XS:' flag were removed and the 5' and 3'-most positions of the remaining read alignments recorded. The imputed fragment densities computed from the ChIP data were then normalized by dividing the density at each position by the fraction of positions within *L* base pairs upstream that were covered by the start site (5' for positive-strand density and 3' for negative-strand density) of a uniquely-mapped genomic read. Positions with 0 uniquely-mappable read start sites within *L* base pairs upstream regions were excluded from further analysis.

In order to compare between species, we transferred data from query organisms to the reference genome using the one-to-one filtered chain files described previously, and calculated the Pearson's correlation between the concatenated chromosome vectors of reference and reference-mapped query data. The evolution of the correlation was modeled and compared using the statistical framework described above.

Genome sequence evolutionary rates (related to Figure 1.2 and Figure 1.5c-d). We defined the cross-species similarity of genomic sequences as the percentage of randomly sampled segments retaining homology. Within the genomes of the reference species, we delineated the boundaries of the regions from which to sample: whole genome (**Figure 1.2; Supplementary Figure S1.4**), intergenic regions in accessible chromatin and intergenic regions in inaccessible chromatin (**Figure 1.5; Supplementary Figure S1.8**). We used the BEDTools shuffle (Quinlan and Hall 2010) to randomize the locations of 5,000 segments of 75 bp length within the delineated boundaries using the option "-noOverlapping." The resulting 5,000 shuffled segments were then mapped across species using the LiftOver tool with minMatch parameter 0.001 (Rosenbloom et al. 2015). We then calculated the percentage of segments that were successfully mapped (*i.e.*, retained homology), excluding segments that mapped to a region longer than 1000bp. The entire simulation was repeated 20 times, starting each time with different sets of 5,000 segments. The percentages of segments retaining homology were recorded for each of the 20 simulations, and averaged for each pair of species. These averages were plotted and used as inputs for our statistical framework. Varying the minMatch parameter of the LiftOver tool to 0.5 and segment length to 150 bp allowed us to verify that the observed trends were robust to sequence similarity thresholds and length sampled (**Supplementary Figure S1.4; Supplementary Figure S1.8)**.

Nucleotide substitution rate within retained genomic segments (related to Supplementary Figure S1.3)**.** The nucleotide sequences of the genomic segments from **Figure 1.2** that retained enough homology to undergo a pairwise alignment were extracted using the getfasta function of BEDtools (Quinlan and Hall 2010). These sequences were then pairwise aligned using EMBOSS suite's implementation of Smith-Waterman local alignment (Rice, Longden, and Bleasby 2000). Default values for gap open penalty (10), gap extend penalty (0.5) and scoring matrix

(EDNAFULL) were used to dynamically choose the best local alignment between reference and query sequences. For each cross-species comparison, we calculated the average percent identity of the ungapped alignments of all the segments across 20 randomizations. This procedure yielded values similar to those described previously for the mouse-human (Mouse Genome Sequencing et al. 2002) and *Drosophila melanogaster – Drosophila pseudoobscura* comparisons (Richards et al. 2005). The average percent identity of ungapped alignments were used as inputs for our statistical framework, revealing that a model that incorporates lineage labels significantly improved fit to the data relative to a common model (P < 0.05; **Supplementary Figure S1.3**).

Motif evolutionary rates (related to Figure 1.5a-b)**.** Using the FIMO tool (Grant, Bailey, and Noble 2011) in the MEME suite (Bailey et al. 2009), the genomes of *D. melanogaster* and *M. musculus domesticus* were scanned for matches to experimentally-determined position-weight matrixes corresponding to the GSTFs of interest. Motif matches were called significant according to the default threshold of FIMO, $P<10^{-4}$. The genomic coordinates of significant motif matches were mapped to the other species within the same lineage using LiftOver (minMatch 0.001). The corresponding coordinates (*Mapped*) were then extended by 50 bp, and the resulting segments were scanned for motif occurrence (*Mappedwithmotif*). In order to estimate background expectation, we randomly shuffled the locations of the *Mapped* segments and scanned these shuffled segments for motifs (*ShuffledMappedwithmotif*). The percentage of motifs retained relative to background was calculated as:

$$F = \frac{Mappedwithmotif - ShuffledMappedwithmotif}{Mapped} * 100$$

The percentages *F* were then used as measures of cross-species similarity to estimate whether a lineage-aware model would describe the evolution of DNA binding motifs better than a common model (**Supplementary Figure S1.7).**

1.6 Figures

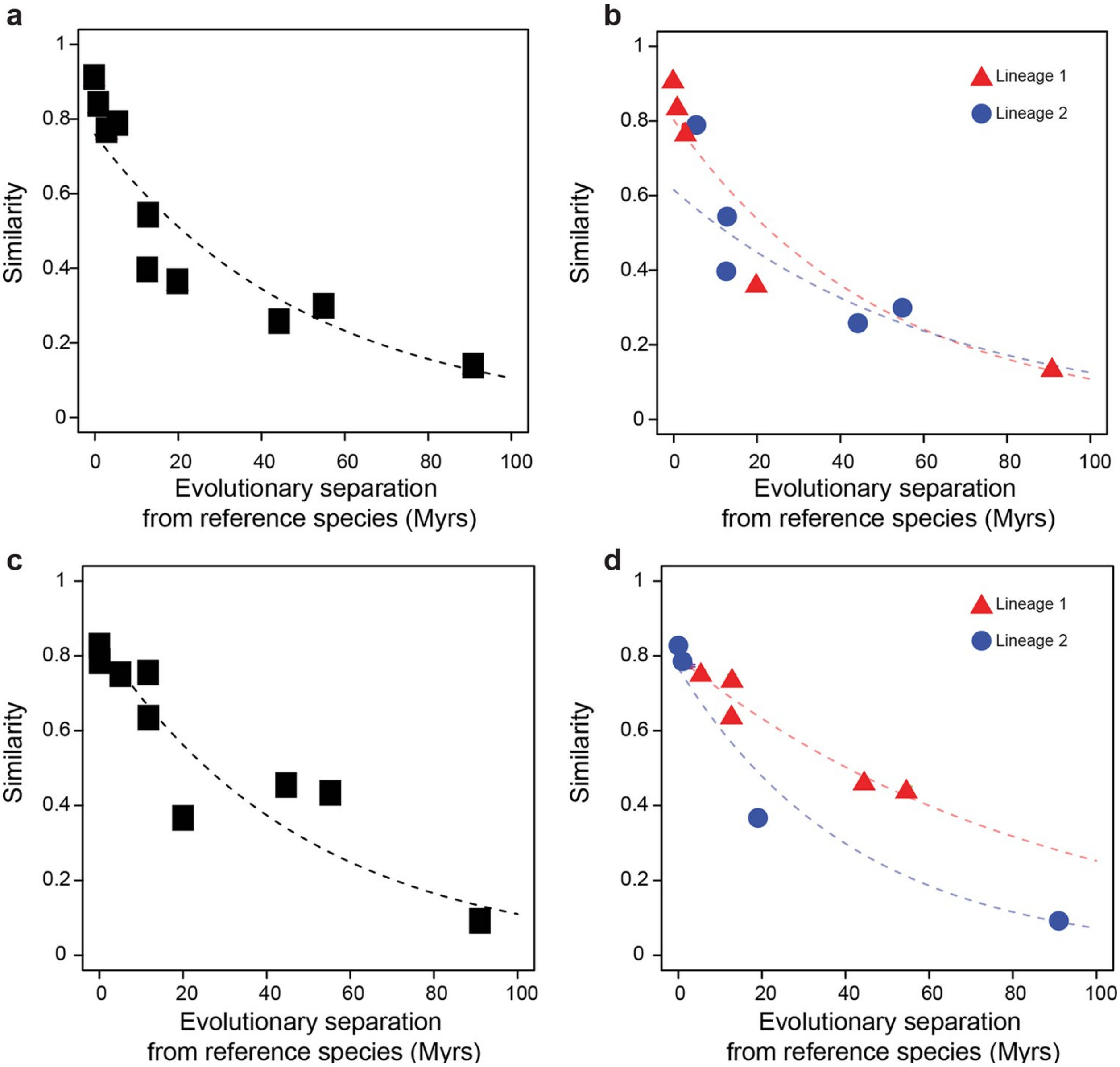


Figure 1.1: Statistical framework to evaluate differences in evolutionary rates of change. Throughout this study we frequently evaluated whether the rate of evolutionary divergence of a given layer of transcriptional regulation differs between lineages. Our approach is equivalent to asking: if the lineage labels were hidden, would one be able to tell that the data points correspond to several lineages or would they seem equally likely to belong to a common distribution? (**a**, **b**) Depict an example of statistically indistinguishable evolutionary rates. Without lineage labels (**a**), the similarity data are modeled by an exponential decay as well as with lineage labels (**b**). Adding lineage labels does not significantly improve the fit. (**c**, **d**) Depict an example of statistically different evolutionary rates. Adding lineage labels (**d**) significantly improves the fit of an exponential decay model over unlabeled data (**c**).

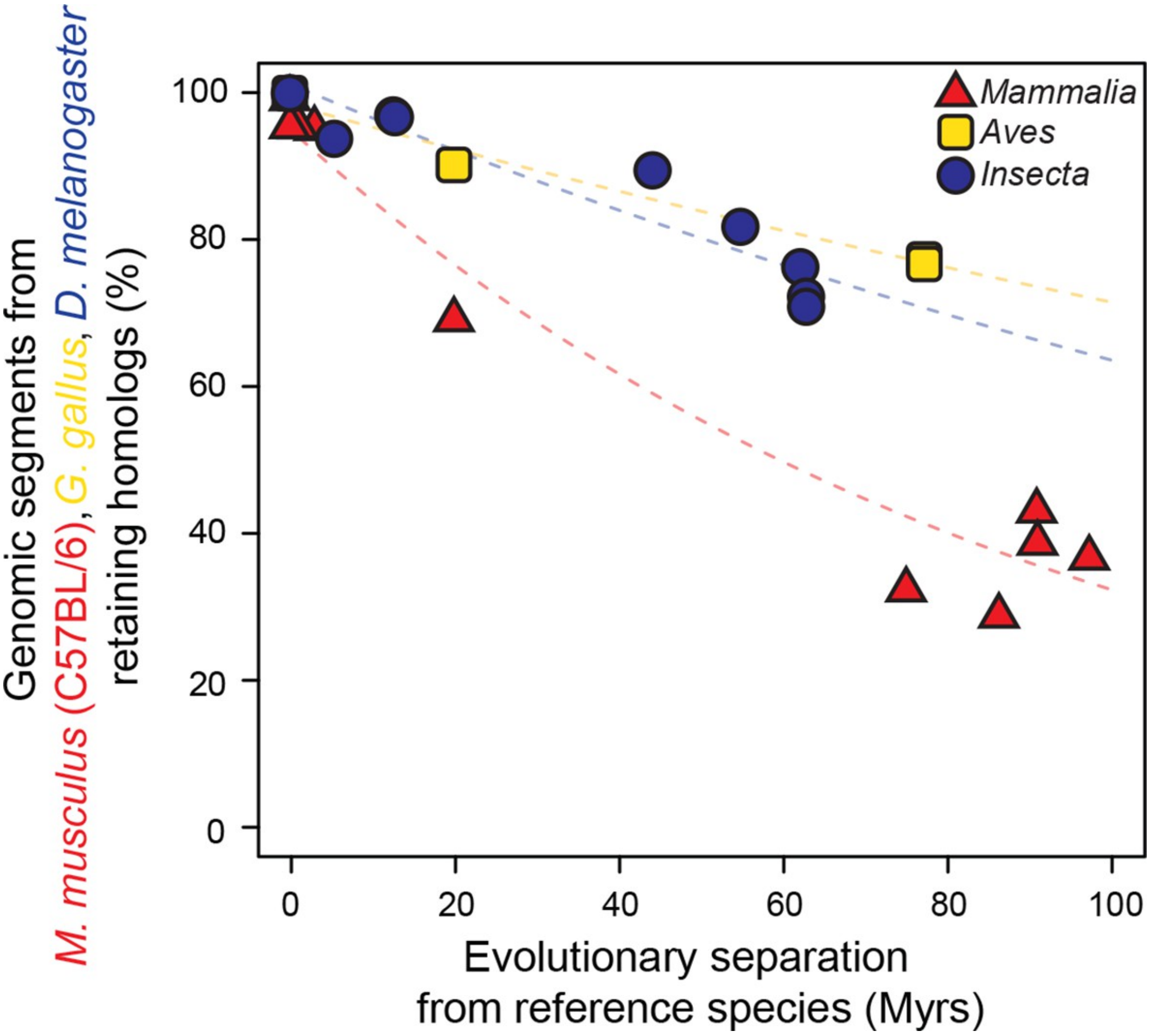


Figure 1.2: Genomic sequences evolve more rapidly in mammals than in birds and insects. The evolutionary retention of 5,000 randomly sampled 75 bp segments was averaged over 20 trials. Organisms compared to reference species are as follows: M. musculus domesticus (AJ), M. musculus castaneus, M. spretus, rat, guinea pig, rabbit, human, chimpanzee and dog for Mammalia; turkey, zebrafinch and flycatcher for Aves; D. simulans, D. erecta, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis, D. willistoni and D. Grimshawi for Insecta. Colored dashed lines: lineage-specific exponential fits, here and in all following displays. The trends were robust to variations in segment length and sequence similarity filters (**Supplementary Figure S1.4**).

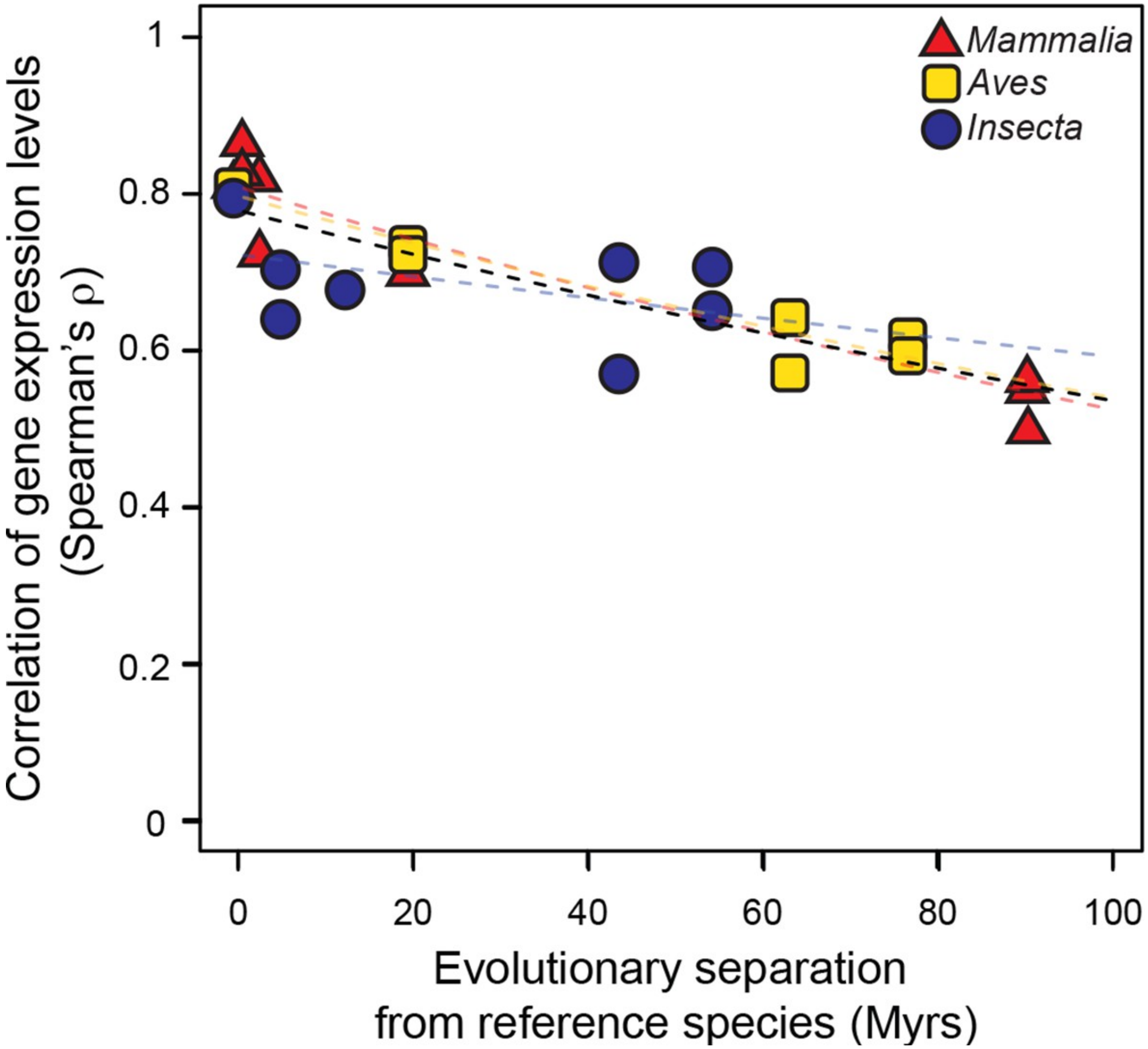
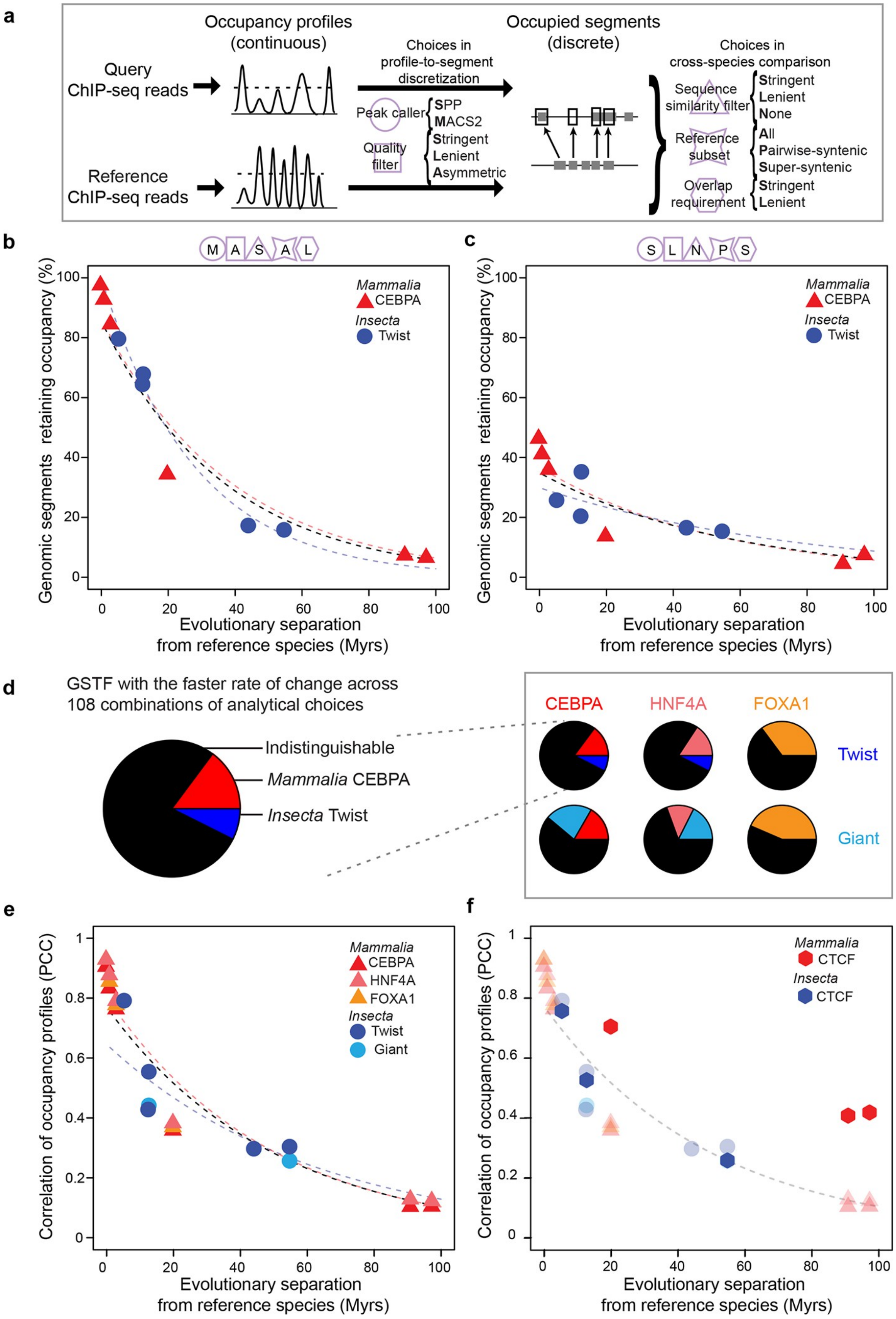


Figure 1.3: Gene expression levels diverge at a common rate in mammals, birds and insects. Gene expression levels were derived independently from two RNA-seq experiments for each reference species and then correlated against each other and against gene expression levels derived from individual experiments in other species within the same lineage. Black dashed line: lineage-naïve exponential fit of all the data, without differentiating the lineages, here and in all following displays. Organisms compared to reference species are as follows: *M. musculus castaneus*, *M. spretus*, rat, human and gorilla for *Mammalia*; turkey, duck and flycatcher for *Aves*; *D. simulans*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura* for *Insecta*.

Figure 1.4: GSTF occupancy diverges at a common rate in mammals and insects.
a, Estimating shared GSTF occupancy across species requires multiple parameter choices. This diagram summarizes the main steps involved in comparing GSTF-occupied segments across species, showing a representative sample of choices at each step (steps represented by purple shapes, specific choices by the first letter bolded). The detailed methods and specific choices illustrated here and implemented in panels **b – d** are described in **Methods**. **b, c,** An example of different analytical choices leading to different results despite starting from the same underlying data. Organisms compared to reference species are as follows: *M. musculus domesticus* (AJ), *M. musculus castaneus*, *M. spretus*, rat, human and dog for *Mammalia*; *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura* for *Insecta*. **d**, Most combinations of choices yield indistinguishable evolutionary rates of GSTF binding patterns across lineages. The comparison of Twist and CEBPA is enlarged to show the color labels corresponding to the statistical interpretation regarding relative evolutionary rates. **e,** A genome-wide comparison of GSTF occupancy profiles at single-nucleotide resolution shows indistinguishable evolutionary rates for CEBPA, HNF4A and FOXA1 in mammals and for Twist and Giant in insects. PCC: Pearson correlation coefficient. **f,** CTCF occupancy is highly conserved in mammals**.** Transparent points and lines are identical as panel **e**. Hexagons correspond to cross-species correlations of CTCF occupancy at single-nucleotide resolution.


a
Occupancy profiles (continuous)
Query ChIP-seq reads
Reference ChIP-seq reads
Choices in profile-to-segment discretization
Peak caller
Quality filter
SPP
MACS2
Stringent
Lenient
Asymmetric
Occupied segments (discrete)
Choices in cross-species comparison
Sequence similarity filter
Reference subset
Overlap requirement
Stringent
Lenient
None
All
Pairwise-syntenic
Super-syntenic
Stringent
Lenient
b
M A S A L
Genomic segments retaining occupancy (%)
Evolutionary separation from reference species (Myrs)
0
20
40
60
80
100
Mammalia
CEBPA
Insecta
Twist
c
S L N P S
Genomic segments retaining occupancy (%)
Evolutionary separation from reference species (Myrs)
Mammalia
CEBPA
Insecta
Twist
d
GSTF with the faster rate of change across 108 combinations of analytical choices
Indistinguishable
Mammalia CEBPA
Insecta Twist
CEBPA
HNF4A
FOXA1
Twist
Giant
e
Correlation of occupancy profiles (PCC)
Evolutionary separation from reference species (Myrs)
0
0.2
0.4
0.6
0.8
1
Mammalia
CEBPA
HNF4A
FOXA1
Insecta
Twist
Giant
f
Correlation of occupancy profiles (PCC)
Evolutionary separation from reference species (Myrs)
Mammalia
CTCF
Insecta
CTCF

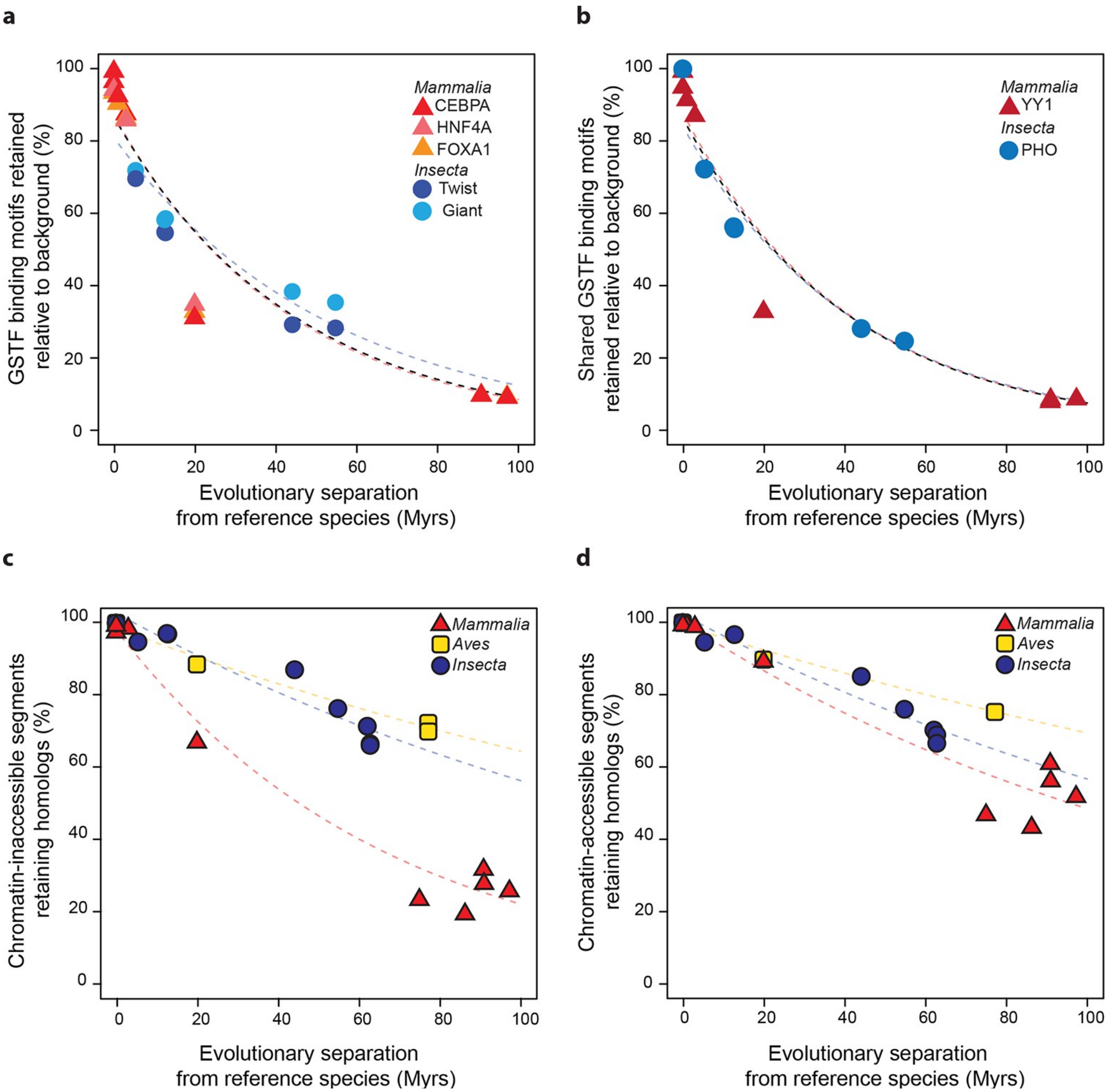
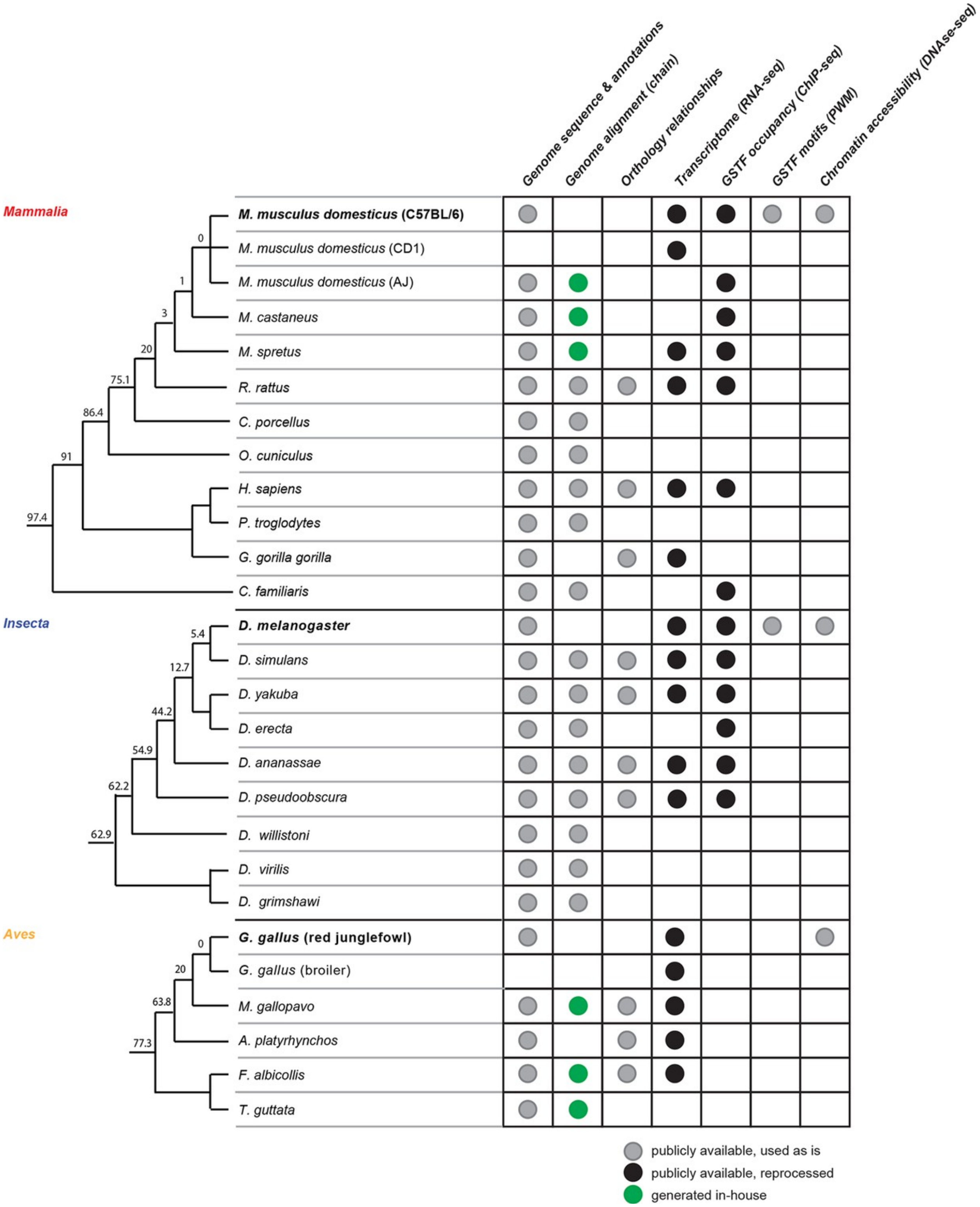


Figure 1.5: Regulatory sequences diverge at similar rates across lineages.
a, The motifs for CEBPA, HNF4A and FOXA1 in mammals and for Twist and Giant in insects are retained at a common rate. Organisms compared to reference species are the same as **Figure 1.4**. **b,** The motifs for GSTFs shared in mammals and insects are retained at common rates. One example is shown here for the motifs corresponding to PHO (FBgn0002521) in *D. melanogaster* and YY1 (ENSMUSG00000021264) in *M. musculus*, which are orthologous GSTFs. Eleven other cases of motif evolution for shared GSTFs conserved in mammals and insects are shown in **Supplementary Figure S1.7**. Organisms compared to reference species are as in **Figure 1.4**. **c, d,** Chromatin-accessible sequences are retained at similar rates in mammals, birds and insects. Analyses were performed as in **Figure 1.2**, limiting sampling to the inaccessible (**c**) and accessible (**d**) portions of the intergenic regions. Organisms compared to reference species are the same as **Figure 1.2**. The trends were robust to variations in segment length and sequence similarity filters (**Supplementary Figure S1.8**).

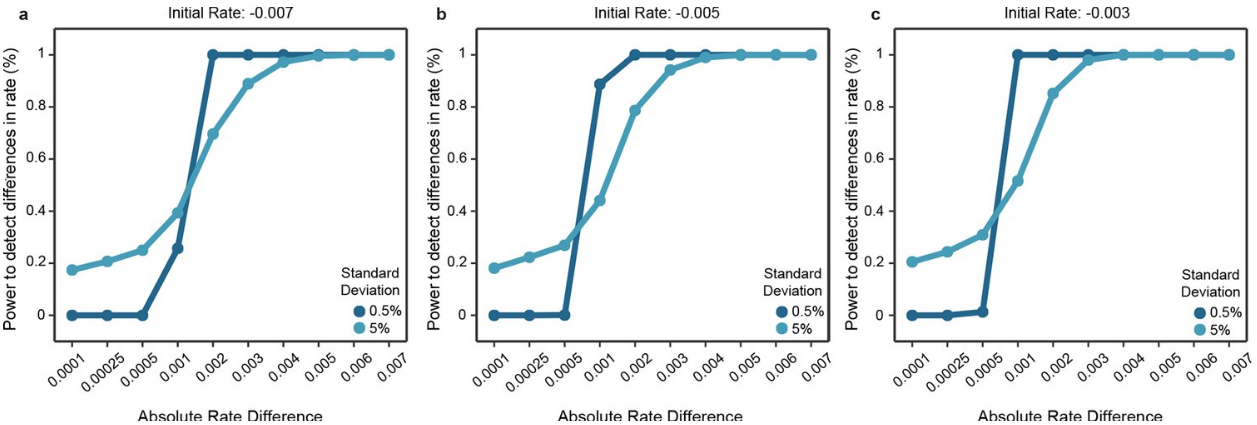
1.7 Supplementary figures

Supplementary Figure S1.1: Comparative genomics platform for studying transcriptional network evolution across three metazoan lineages.
The phylogenetic trees indicate the evolutionary relationships between the organisms included in this study. The trees are not drawn to scale. The numbers at each branch split represent the evolutionary distance in Myrs separating the organisms at the end of the lower branch from the reference species, whose names are bolded.



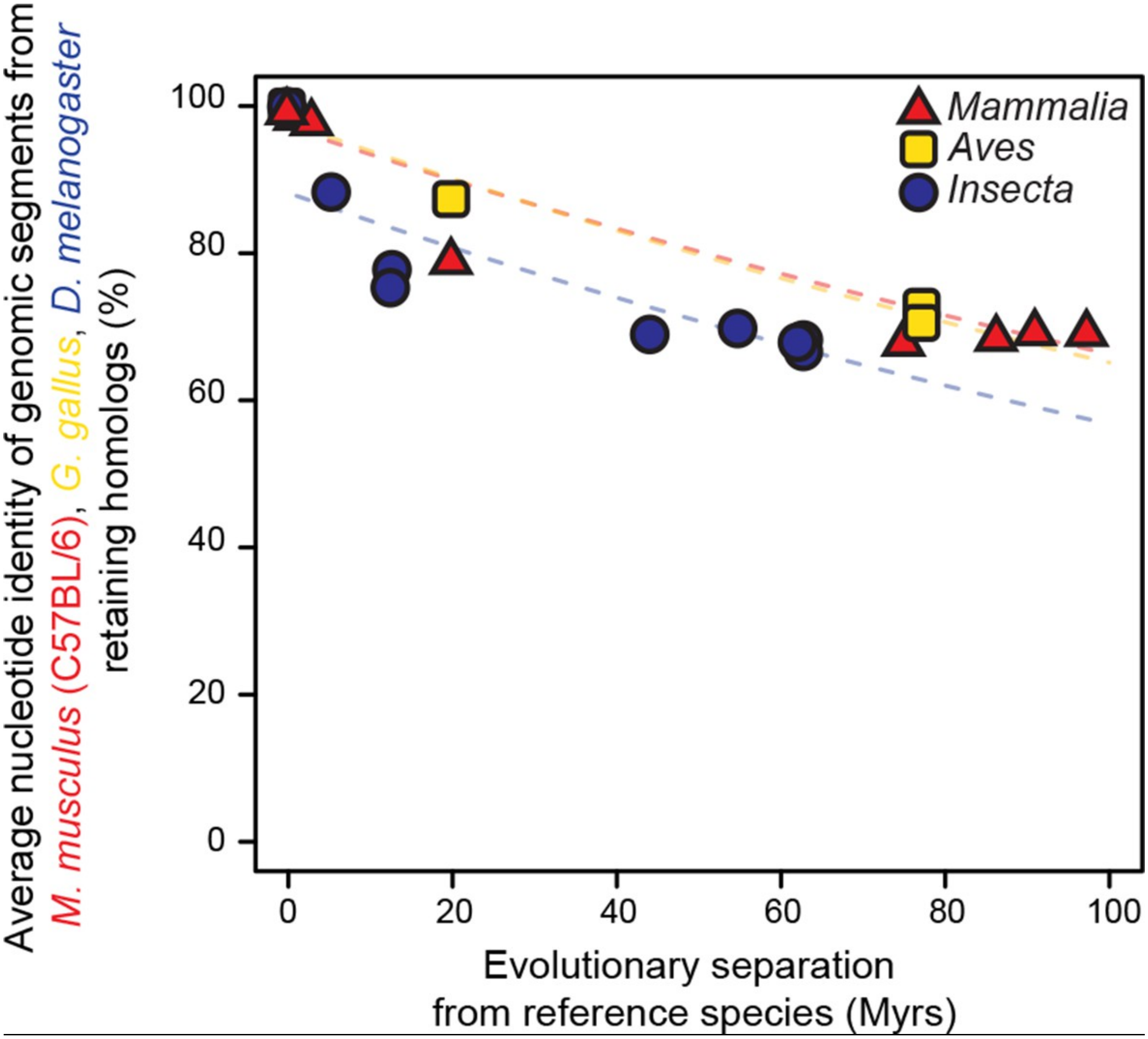
	Genome sequence & annotations	*Genome alignment (chain)*	*Orthology relationships*	*Transcriptome (RNA-seq)*	*GSTF occupancy (ChIP-seq)*	*GSTF motifs (PWM)*	*Chromatin accessibility (DNAse-seq)*
***M. musculus domesticus* (C57BL/6)**	○			●	●	○	○
M. musculus domesticus (CD1)				●			
M. musculus domesticus (AJ)	○	🟢			●		
M. castaneus	○	🟢			●		
M. spretus	○	🟢		●	●		
R. rattus	○	○	○	●	●		
C. porcellus	○	○					
O. cuniculus	○	○					
H. sapiens	○	○	○	●	●		
P. troglodytes	○	○					
G. gorilla gorilla	○		○	●			
C. familiaris	○	○			●		
D. melanogaster	○			●	●	○	○
D. simulans	○	○	○	●	●		
D. yakuba	○	○	○	●	●		
D. erecta	○	○			●		
D. ananassae	○	○	○	●	●		
D. pseudoobscura	○	○	○	●	●		
D. willistoni	○	○					
D. virilis	○	○					
D. grimshawi	○	○					
***G. gallus* (red junglefowl)**	○			●			○
G. gallus (broiler)				●			
M. gallopavo	○	🟢	○	●			
A. platyrhynchos	○		○	●			
F. albicollis	○	🟢	○	●			
T. guttata	○	🟢					

○ publicly available, used as is
● publicly available, reprocessed
🟢 generated in-house

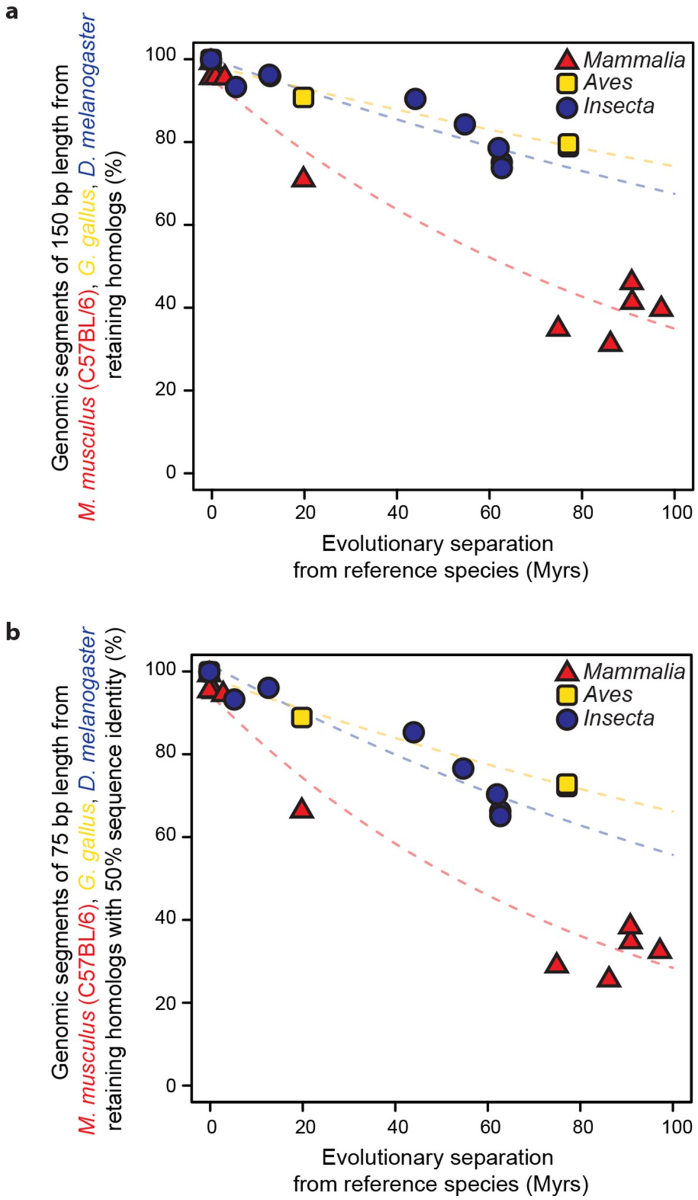


Supplementary Figure S1.2: Power of the statistical framework to evaluate differences in evolutionary rates.

a-c, Depict the sensitivity of our statistical framework to detect differences in 1000 simulations. The initial rates of one clade was fixed to either -0.007 **a**, -0.005 **b**, or -0.003 **c**, and data were simulated by modeling an exponential decay where samples were drawn from a Gaussian distribution with standard deviation fixed to 0.5% or 5%. The second clade's rate was modeled according to the absolute difference in rates with steps shown in the x axis and sampled similarly as for the first clade. Simulated data were used as input to our statistical framework and the frequency of detecting a significant difference is shown.

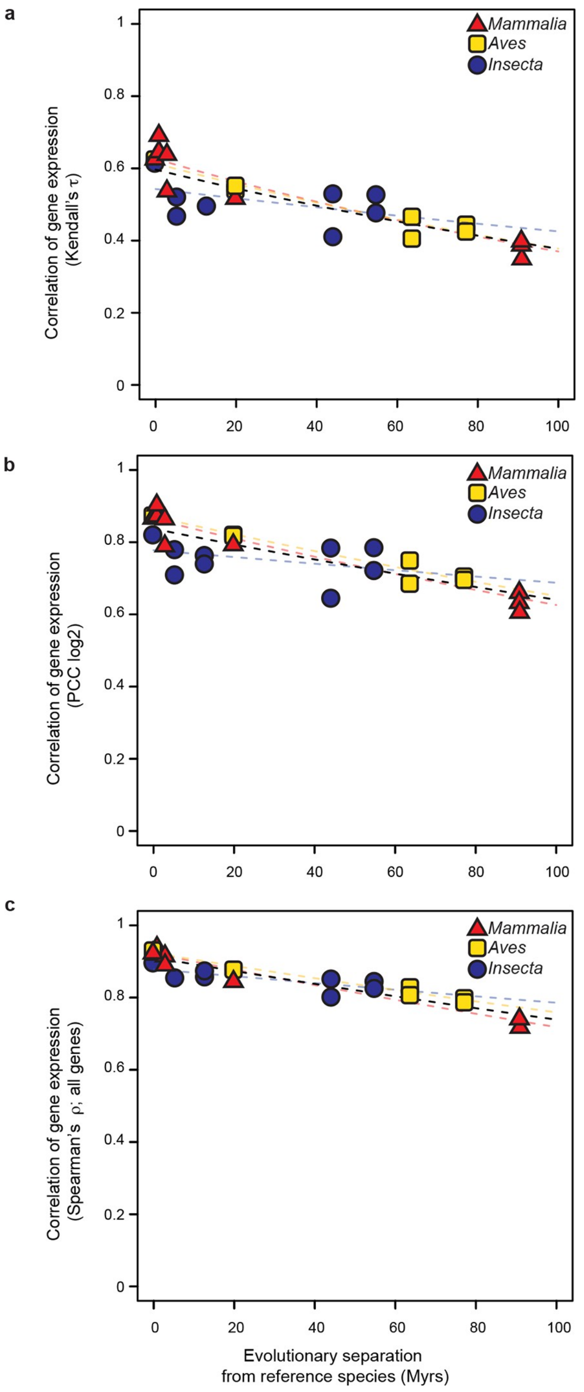


Supplementary Figure S1.3: Genomic segments retaining homologs are highly conserved at the nucleotide level.
The genomic segments found to retain homologs in **Figure 1.2** were aligned to their homologous regions. The average nucleotide identity of the corresponding ungapped alignment is shown here. Evolutionary rates are slightly but significantly different among lineages ($P < 0.05$).

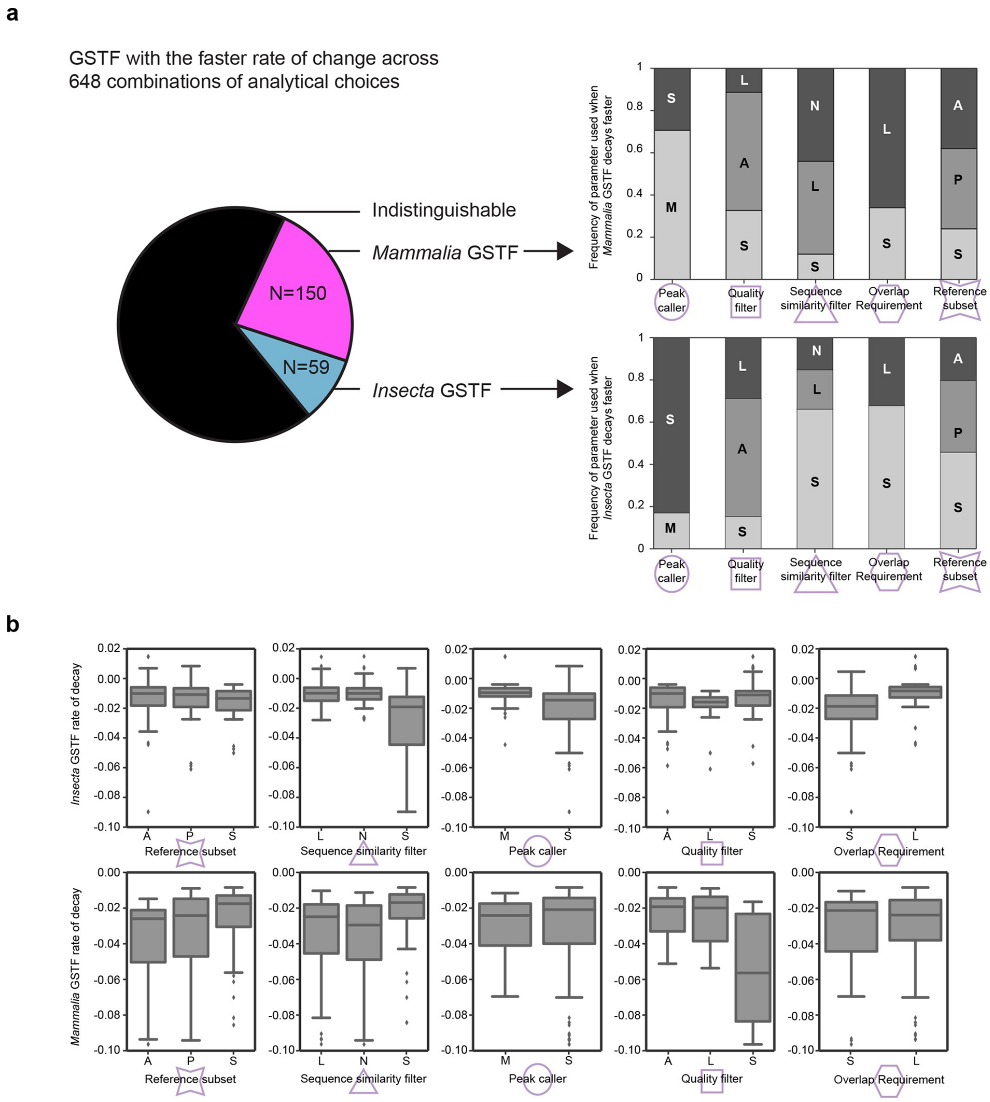


Supplementary Figure S1.4: Retention of genomic segments is robust to changes in sampled region size and sequence identity threshold.

The evolutionary retention of 5000 randomly sampled 75 bp segments was averaged over 20 trials. Organisms compared to reference species are as follows: M. musculus domesticus (AJ), M. musculus castaneus, M. spretus, rat, guinea pig, rabbit, human, chimpanzee and dog for Mammalia; turkey, zebrafinch and flycatcher for Aves; D. simulans, D. erecta, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis, D. willistoni and D. grimshawi for Insecta. Colored dashed lines: lineage-specific exponential fits, here and in all following displays. The trends were robust to variations in segment length and sequence similarity filters.

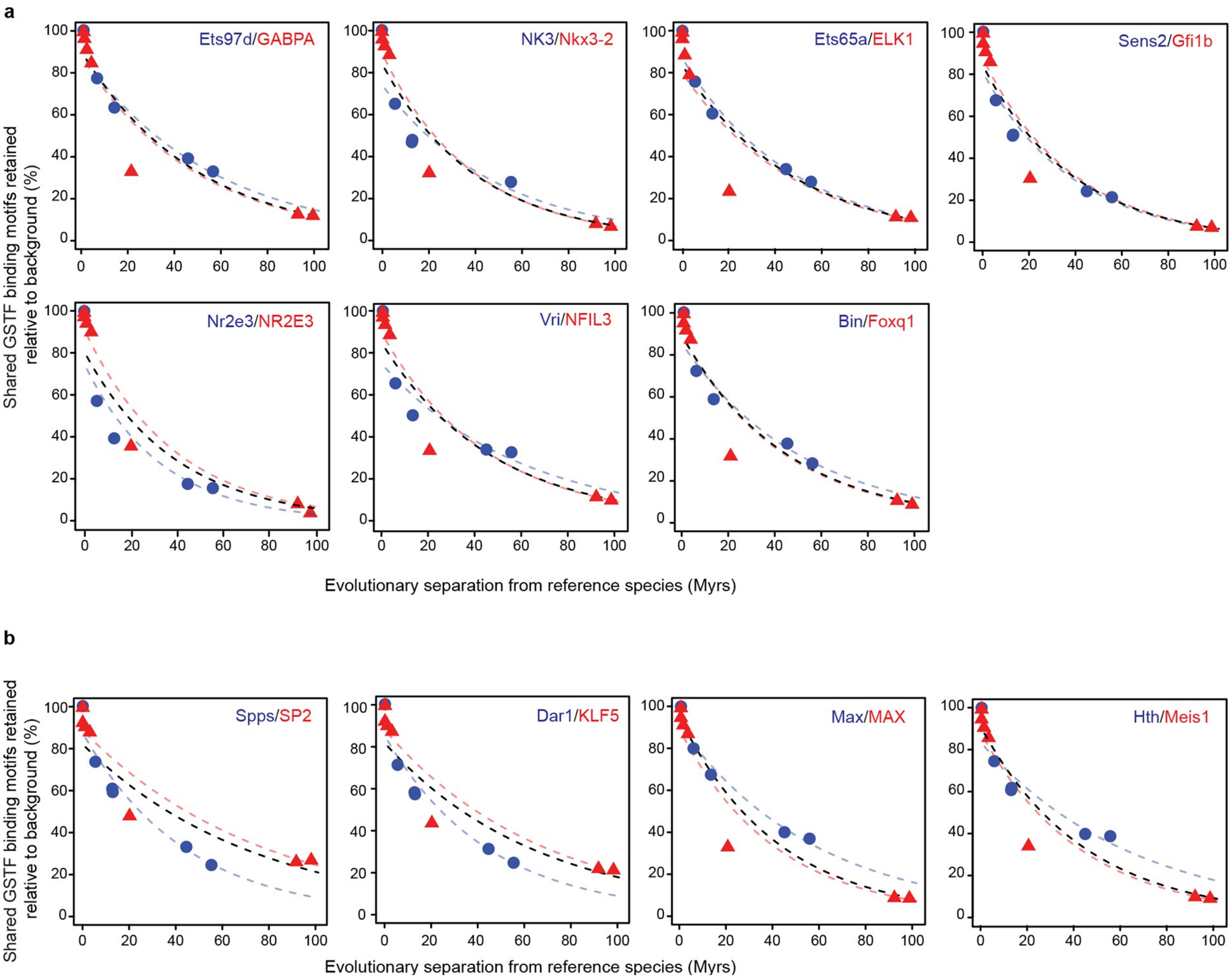


Supplementary Figure S1.5: The common evolutionary rate of gene expression levels presented in Figure 1.3 is robust to changes in correlation metrics or expression threshold. **a-b**, The gene expression levels used in **Figure 1.3** were correlated with alternative correlation metrics, Kendalls τ (**a**) and Pearson's r (**b**). The resulting evolutionary rates remained statistically indistinguishable. (**c**) The gene expression level of all genes were analyzed rather than excluding the values below 5 TPM as was done in **Figure 1.3**. The resulting evolutionary rates remained statistically indistinguishable.



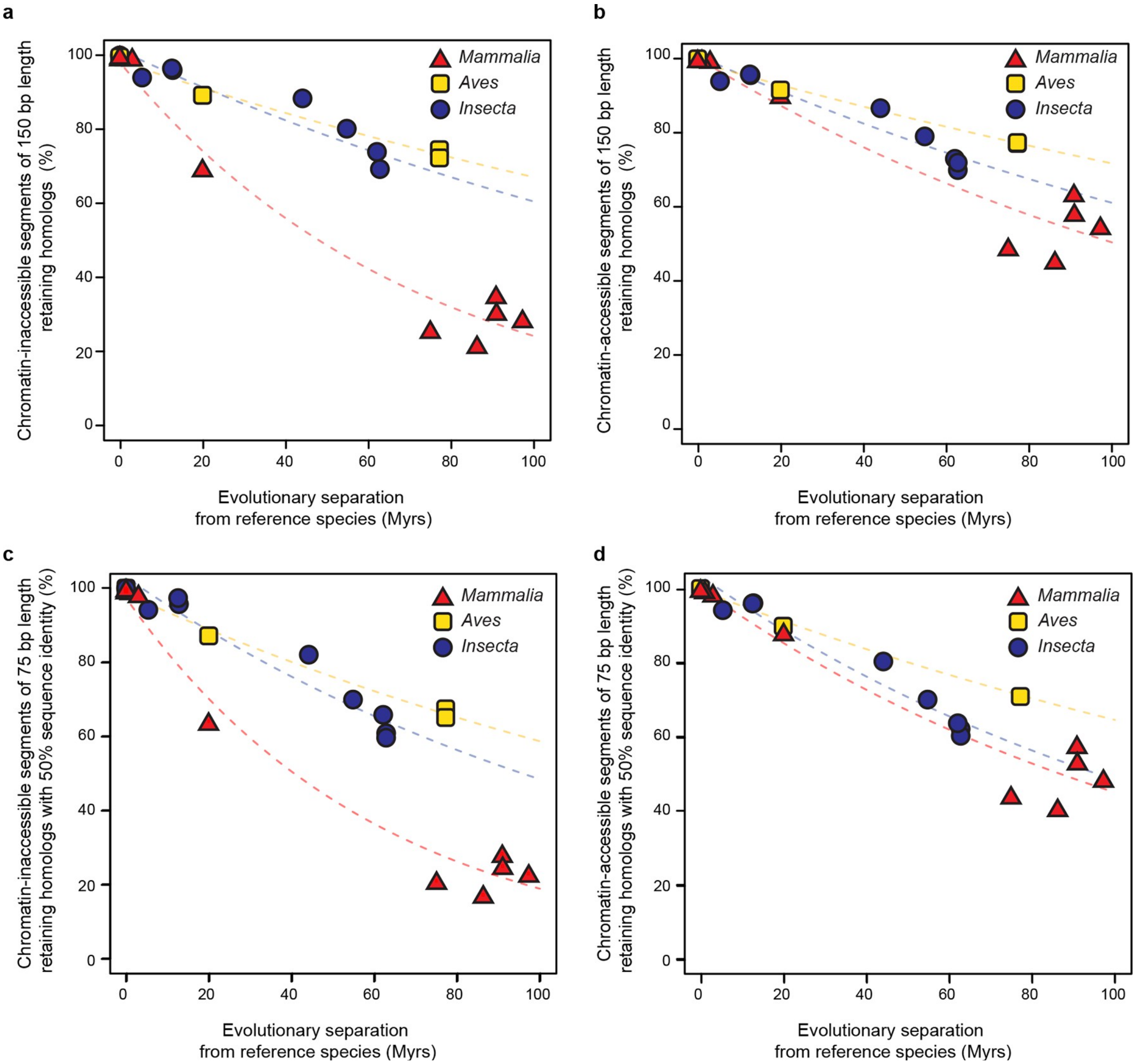
Supplementary Figure S1.6: Measured GSTF binding divergence rates are influenced by parameter choices.

(**a**) The pie chart on the left shows the frequency at which either the mammalian or insect GSTF was found to evolve faster for 648 comparisons using different combinations of analytical choices. The majority of comparisons showed indistinguishable rates. The stacked histograms indicate how often a parameter was used when a difference in divergence rate was detected. For instance, 106/150 cases where Mammalia factors decayed significantly faster used MACS2 as a peak caller , whereas 49/59 cases of Insecta GSTF decaying faster used SPP. Interestingly, asymmetric quality filters showed an enrichment for Mammalia GSTFs decaying faster (84/150) as well as for Insecta GSTFs decaying faster (33/59). (**b**) Boxplots showing general influence of parameter choices on individual decay rates of Insecta (top) and Mammalia (bottom). Only instances when a significant fit was detected are considered. For example, for mammalian GSTFs, stringent quality filters yielded slightly faster decay rates than asymmetric or lenient quality filters. Summary of all parameter choices and the results are shown in **Supplementary File S1.2**. These parameters are further elaborated in Figure 1.4 and **Supplementary File S1.4**.



Supplementary Figure S1.7: Conservation of cis-regulatory motifs for GSTFs conserved across insects and mammals.

(**a**) Seven shared GSTFs whose motifs are retained at indistinguishable rates in mammals and insects. The evolution of these motifs behave similarly to that of the example shown in **Figure 1.5**. (**b**) Four conserved GSTFs whose motifs are retained at different rates in mammals and insects.



Supplementary Figure S1.8: Retention of intergenic genomic segments in accessible- and inaccessible-chromatin is robust to changes in sampled region size and sequence identity threshold.

(**a, b**) Repeating the procedure used in Figure 1.5c and Figure 1.5d sampling segments of different length (150 bp) and (**c**, **d**) increasing the LiftOver minMatch (0.5).

1.8 Supplementary tables & files

Supplementary Table S1.1: Published ChIP-seq studies comparing binding locations of GSTFs in closely related metazoans used different technical methodologies to estimate divergence rates.

This table lists variations of five parameters used to identify occupied segments and compare their locations across species by the original studies that generated the ChIP-seq data we have reprocessed in our study. We only summarize this set of five parameters (peak caller, quality filter, sequence similarity filter, reference subset, overlap requirement) because those are the focus of our re-analyses (**Figure 1.4**). For more details on the specific method, please see the original study.

Lineage	GSTF	Peak caller	Quality filter	Sequence similarity filter	Reference subset	Overlap requirement	Study
Insects	GIANT	MACS	Relative	None	Pairwise-synteny	Stringent	Bradley *et al. PLoS Biology*, 2010
Insects	TWIST	MACS	Asymmetric	Stringent	All	Stringent	He *et al. Nature Genetics*, 2011
Insects	GIANT	MACS and Grizzly Peak	Lenient	None	Multiple choices	Lenient	Paris *et al. PLoS Biology*, 2013
Mammals	CEBPA, HNF4A	SWEMBL	Stringent	None	Super-synteny	Lenient	Schmidt *et al. Science*, 2010
Mammals	CEBPA, HNF4A, FOXA1	SWEMBL	Stringent	None	All	Lenient	Stefflova *et al. Cell*, 2013

Supplementary Table S1.2 Accession numbers used in RNA-seq analyses.

Species	Accession*
M. musculus domesticus (C57BL/6)	ERR185942
M. musculus domesticus (CD1)	SRR908295
G. gallus (red junglefowl)	SRR306720
G. gallus (broiler)	SRR998879
D. melanogaster	SRR166808
D. melanogaster	SRR030231
M. castaneus	ERR120692
M. spretus	ERR476403
R. norvegicus	SRR1178064
H. sapiens	ENCFF283RUU, ENCFF187OKV
G. gorilla	SRR306809
D. simulans	SRR166812
D. yakuba	SRR166821
D. ananassae	SRR166825
D. pseudoobscura	SRR166829
M. gallopavo	SRR1334841
A. platyrhynchos	SRR064720
F. albicollis	ERR168726,ERR168727,ERR168728,ERR168729,ERR168730

*',' across multiple accession numbers indicates that these accession numbers were concatenated to form a single sample, otherwise only one accession number was used as a sample

Supplementary Table S1.3: Parameters used to build chain files among vertebrate genomes
The table depicts the order in which tools were used to generate chain files (Methods) along with their specific parameters (starting with lastZ). We further note that we generated pairwise alignments for mice using chromosomes that were named the same. For *Aves*, we generated all pairwise alignments of designated *G. gallus* reference chromosomes with all scaffolds in the other *Aves* species

Tool	Mice parameters	Birds parameters	Description
lastZ	--notransition --nogapped --format=lav --hspthresh=6000 --step=100 --progress	--transition=1 --format=lav --hspthresh=3000 --gappedthresh=3000 --inner=2000 --masking=50	Generates pairwise alignments between chromosomes
lavToPsl	N/A	N/A	Converts lav format to psl format
axtChain	-linearGap=loose -psl	-linearGap=medium -minScore=3000 -psl	Chains alignments together
chainMergeSort | chainSplit	-lump=50	-lump=50	Combines sorted files into larger sorted files, and splits the chains by target/query sequence
chainNet	default	default	Makes alignment nets out of chains
netChainSubset	default	default	Creates a chain file with subset of chains that appear in the net
Custom python script to filter chains	chainscoremin=5000000000	not done	Filters chains to keep the best chain, one per chromosome -- only done with mouse alignments

Supplementary File S1.1: Accession numbers used in ChIP-seq analyses
Data shows the results of each analytical pipeline, and whether the fit was considered significant or insignificant.

Supplementary File S1.2: 648 segment-based ChIP analyses
Data shows the results of each analytical pipeline, and whether the fit was considered significant or insignificant.

Supplementary File S1.3: Influence of parameters choices when assessing GSTF binding divergence at segment resolution in mammals and insects.
This file contains text that further elaborates the results observed when assessing various analytical frameworks to analyze transcription factor evolution.

1.9 Author contributions

A-RC, TW, DS contributed equally to the formation of this work. A-RC, TW, DS, JFK, AY, JC, and TI conception and design, Analysis and interpretation of data, drafting and revising the article.

1.10 Acknowledgements

This work was supported by the National Institutes of Health: R01 GM084279 and P50 GM085764 awarded to T.I., T32 GM008666 awarded to T.W. and the Pathway to Independence K99 GM108865 awarded to A-R.C. The authors are grateful to Drs. Pollard K.S., Wittkopp P.J., Burke M., Charloteaux B. and Coolon J.D. for review of the manuscript prior to submission and to the editors and reviewers for their help in improving the manuscript after submission.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Evidence for a common evolutionary rate in metazoan transcriptional networks" in *eLife*, 2015 by Anne-Ruxandra Carvunis, Tina Wang, Dylan Skola, Alice Yu, Jonathan Chen, Jason F Kreisberg and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

1.11 References

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. "MEME SUITE: tools for motif discovery and searching." *Nucleic Acids Res* 37 (Web Server issue):W202-8. doi: 10.1093/nar/gkp335.

Bardet, A. F., Q. He, J. Zeitlinger, and A. Stark. 2012. "A computational pipeline for comparative ChIP-seq analyses." *Nat Protoc* 7 (1):45-61. doi: 10.1038/nprot.2011.420.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15):2114-20. doi: 10.1093/bioinformatics/btu170.

Bradley, R. K., X. Y. Li, C. Trapnell, S. Davidson, L. Pachter, H. C. Chu, L. A. Tonkin, M. D. Biggin, and M. B. Eisen. 2010. "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species." *PLoS Biol* 8 (3):e1000343. doi: 10.1371/journal.pbio.1000343.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grutzner, S. Bergmann, R. Nielsen, S. Paabo, and H. Kaessmann. 2011. "The evolution of gene expression levels in mammalian organs." *Nature* 478 (7369):343-8. doi: 10.1038/nature10532.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. "Genetic dissection of transcriptional regulation in budding yeast." *Science* 296 (5568):752-5. doi: 10.1126/science.1069516.

Chan, E. T., G. T. Quon, G. Chua, T. Babak, M. Trochesset, R. A. Zirngibl, J. Aubin, M. J. Ratcliffe, A. Wilde, M. Brudno, Q. D. Morris, and T. R. Hughes. 2009. "Conservation of core gene expression in vertebrate tissues." *J Biol* 8 (3):33. doi: 10.1186/jbiol130.

Chen, Z. X., D. Sturgill, J. Qu, H. Jiang, S. Park, N. Boley, A. M. Suzuki, A. R. Fletcher, D. C. Plachetzki, P. C. FitzGerald, C. G. Artieri, J. Atallah, O. Barmina, J. B. Brown, K. P. Blankenburg, E. Clough, A. Dasgupta, S. Gubbala, Y. Han, J. C. Jayaseelan, D. Kalra, Y. A. Kim, C. L. Kovar, S. L. Lee, M. Li, J. D. Malley, J. H. Malone, T. Mathew, N. R. Mattiuzzo, M. Munidasa, D. M. Muzny, F. Ongeri, L. Perales, T. M. Przytycka, L. L. Pu, G. Robinson, R. L. Thornton, N. Saada, S. E. Scherer, H. E. Smith, C. Vinson, C. B. Warner, K. C. Worley, Y. Q. Wu, X. Zou, P. Cherbas, M. Kellis, M. B. Eisen, F. Piano, K. Kionte, D. H. Fitch, P. W. Sternberg, A. D. Cutter, M. O. Duff, R. A. Hoskins, B. R. Graveley, R. A. Gibbs, P. J. Bickel, A. Kopp, P. Carninci, S. E. Celniker, B. Oliver, and S. Richards. 2014. "Comparative validation of the D. melanogaster modENCODE transcriptome annotation." *Genome Res* 24 (7):1209-23. doi: 10.1101/gr.159384.113.

Coble, D. J., D. Fleming, M. E. Persia, C. M. Ashwell, M. F. Rothschild, C. J. Schmidt, and S. J. Lamont. 2014. "RNA-seq analysis of broiler liver transcriptome reveals novel responses to high ambient temperature." *BMC Genomics* 15:1084. doi: 10.1186/1471-2164-15-1084.

Consortium, Encode Project. 2012. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489 (7414):57-74. doi: 10.1038/nature11247.

Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. 2009. "Mapping complex disease traits with global gene expression." *Nat Rev Genet* 10 (3):184-94. doi: 10.1038/nrg2537.

Coolon, J. D., C. J. McManus, K. R. Stevenson, B. R. Graveley, and P. J. Wittkopp. 2014. "Tempo and mode of regulatory evolution in Drosophila." *Genome Res* 24 (5):797-808. doi: 10.1101/gr.163014.113.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. 2015. "Ensembl 2015." *Nucleic Acids Res* 43 (Database issue):D662-9. doi: 10.1093/nar/gku1010.

dos Santos, G., A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, W. M. Gelbart, and Consortium FlyBase. 2015. "FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations." *Nucleic Acids Res* 43 (Database issue):D690-7. doi: 10.1093/nar/gku1099.

Goncalves, A., S. Leigh-Brown, D. Thybert, K. Stefflova, E. Turro, P. Flicek, A. Brazma, D. T. Odom, and J. C. Marioni. 2012. "Extensive compensatory cis-trans regulation in the evolution of mouse gene expression." *Genome Res* 22 (12):2376-84. doi: 10.1101/gr.142281.112.

Gong, B., C. Wang, Z. Su, H. Hong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, S. S. Auerbach, W. Tong, and J. Xu. 2014. "Transcriptomic profiling of rat liver samples in a comprehensive study design by RNA-Seq." *Sci Data* 1:140021. doi: 10.1038/sdata.2014.21.

Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. "FIMO: scanning for occurrences of a given motif." *Bioinformatics* 27 (7):1017-8. doi: 10.1093/bioinformatics/btr064.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. 2013. "On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE." *Genome Biol Evol* 5 (3):578-90. doi: 10.1093/gbe/evt028.

He, B. Z., A. K. Holloway, S. J. Maerkl, and M. Kreitman. 2011. "Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules." *PLoS Genet* 7 (4):e1002053. doi: 10.1371/journal.pgen.1002053.

He, Q., A. F. Bardet, B. Patton, J. Purvis, J. Johnston, A. Paulson, M. Gogol, A. Stark, and J. Zeitlinger. 2011. "High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species." *Nat Genet* 43 (5):414-20. doi: 10.1038/ng.808.

He, Y., J. A. Carrillo, J. Luo, Y. Ding, F. Tian, I. Davidson, and J. Song. 2014. "Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells." *Front Genet* 5:308. doi: 10.3389/fgene.2014.00308.

Hedges, S. B. 2009. *The Timetree of Life*: Oxford University Press.

Heinz, S., C. E. Romanoski, C. Benner, K. A. Allison, M. U. Kaikkonen, L. D. Orozco, and C. K. Glass. 2013. "Effect of natural genetic variation on enhancer selection and function." *Nature* 503 (7477):487-92. doi: 10.1038/nature12615.

Hesselberth, J. R., X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. 2009. "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." *Nat Methods* 6 (4):283-9. doi: 10.1038/nmeth.1313.

Huang, Y., Y. Li, D. W. Burt, H. Chen, Y. Zhang, W. Qian, H. Kim, S. Gan, Y. Zhao, J. Li, K. Yi, H. Feng, P. Zhu, B. Li, Q. Liu, S. Fairley, K. E. Magor, Z. Du, X. Hu, L. Goodman, H. Tafer, A. Vignal, T. Lee, K. W. Kim, Z. Sheng, Y. An, S. Searle, J. Herrero, M. A. Groenen, R. P. Crooijmans, T. Faraut, Q. Cai, R. G. Webster, J. R. Aldridge, W. C. Warren, S. Bartschat, S. Kehr, M. Marz, P. F. Stadler, J. Smith, R. H. Kraus, Y. Zhao, L. Ren, J. Fei, M. Morisson, P. Kaiser, D. K. Griffin, M. Rao, F. Pitel, J. Wang, and N. Li. 2013. "The duck genome and transcriptome provide insight into an avian influenza virus reservoir species." *Nat Genet* 45 (7):776-83. doi: 10.1038/ng.2657.

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M. Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder. 2010. "Variation in transcription factor binding among humans." *Science* 328 (5975):232-5. doi: 10.1126/science.1183621.

Kasowski, M., S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, Y. Liu, A. P. Boyle, Q. C. Zhang, F. Zakharia, D. V. Spacek, J. Li, D. Xie, A. Olarerin-George, L. M. Steinmetz, J. B. Hogenesch, M. Kellis, S. Batzoglou, and M. Snyder. 2013. "Extensive variation in chromatin states across humans." *Science* 342 (6159):750-2. doi: 10.1126/science.1242510.

Kellis, M., B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White, and R. C. Hardison. 2014. "Defining functional DNA elements in the human genome." *Proc Natl Acad Sci U S A* 111 (17):6131-8. doi: 10.1073/pnas.1318948111.

Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park. 2008. "Design and analysis of ChIP-seq experiments for DNA-binding proteins." *Nat Biotechnol* 26 (12):1351-9. doi: 10.1038/nbt.1508.

Kumar, S. 2005. "Molecular clocks: four decades of evolution." *Nat Rev Genet* 6 (8):654-62. doi: 10.1038/nrg1659.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor,

G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. 2012. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome Res* 22 (9):1813-31. doi: 10.1101/gr.136184.111.

Langmead, B., and S. L. Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nat Methods* 9 (4):357-9. doi: 10.1038/nmeth.1923.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16):2078-9. doi: 10.1093/bioinformatics/btp352.

Lin, S., Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder. 2014. "Comparison of the transcriptional landscapes between human and mouse tissues." *Proc Natl Acad Sci U S A* 111 (48):17224-9. doi: 10.1073/pnas.1413624111.

Lu, L., Y. Chen, Z. Wang, X. Li, W. Chen, Z. Tao, J. Shen, Y. Tian, D. Wang, G. Li, L. Chen, F. Chen, D. Fang, L. Yu, Y. Sun, Y. Ma, J. Li, and J. Wang. 2015. "The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver." *Genome Biol* 16:89. doi: 10.1186/s13059-015-0652-y.

Mathelier, A., X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. 2014. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles." *Nucleic Acids Res* 42 (Database issue):D142-7. doi: 10.1093/nar/gkt997.

McVicker, G., B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard. 2013. "Identification of genetic variants that affect histone modifications in human cells." *Science* 342 (6159):747-9. doi: 10.1126/science.1242429.

Monson, M. S., R. E. Settlage, K. W. McMahon, K. M. Mendoza, S. Rawal, H. S. El-Nezami, R. A. Coulombe, and K. M. Reed. 2014. "Response of the hepatic transcriptome to aflatoxin B1 in domestic turkey (Meleagris gallopavo)." *PLoS One* 9 (6):e100930. doi: 10.1371/journal.pone.0100930.

Mouse Genome Sequencing, Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T.

Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander. 2002. "Initial sequencing and comparative analysis of the mouse genome." *Nature* 420 (6915):520-62. doi: 10.1038/nature01262.

Ni, X., Y. E. Zhang, N. Negre, S. Chen, M. Long, and K. P. White. 2012. "Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome." *PLoS Biol* 10 (11):e1001420. doi: 10.1371/journal.pbio.1001420.

Niu, D. K., and L. Jiang. 2013. "Can ENCODE tell us how much junk DNA we carry in our genome?" *Biochem Biophys Res Commun* 430 (4):1340-3. doi: 10.1016/j.bbrc.2012.12.074.

Ohlsson, R., V. Lobanenkov, and E. Klenova. 2010. "Does CTCF mediate between nuclear organization and gene expression?" *Bioessays* 32 (1):37-50. doi: 10.1002/bies.200900118.

Ohta, T. 1992. "The Nearly Neutral Theory of Molecular Evolution." *Annual Review of Ecology and Systematics* 23:263-286.

Paris, M., T. Kaplan, X. Y. Li, J. E. Villalta, S. E. Lott, and M. B. Eisen. 2013. "Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression." *PLoS Genet* 9 (9):e1003748. doi: 10.1371/journal.pgen.1003748.

Patro, R., S. M. Mount, and C. Kingsford. 2014. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms." *Nat Biotechnol* 32 (5):462-4. doi: 10.1038/nbt.2862.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6):841-2. doi: 10.1093/bioinformatics/btq033.

R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rebollo, R., M. T. Romanish, and D. L. Mager. 2012. "Transposable elements: an abundant and natural source of regulatory sequences for host genes." *Annu Rev Genet* 46:21-42. doi: 10.1146/annurev-genet-110711-155621.

Rice, P., I. Longden, and A. Bleasby. 2000. "EMBOSS: the European Molecular Biology Open Software Suite." *Trends Genet* 16 (6):276-7.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, B. B. Matthews, M. A. Crosby, A. J. Schroeder, D. Ortiz-Barrientos, C. M. Rives, M. L. Metzker, D. M. Muzny, G. Scott, D. Steffen, D. A. Wheeler, K. C. Worley, P. Havlak, K. J. Durbin, A. Egan, R. Gill, J. Hume, M. B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K. P. Clerc-Blankenburg, I. Dubchak, M. A. Noor, W. Anderson, K. P. White, A. G. Clark, S. W. Schaeffer, W. Gelbart, G. M. Weinstock, and R. A. Gibbs. 2005. "Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution." *Genome Res* 15 (1):1-18. doi: 10.1101/gr.3059305.

Rosenbloom, K. R., J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. Smit, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. 2015. "The UCSC Genome Browser database: 2015 update." *Nucleic Acids Res* 43 (Database issue):D670-81. doi: 10.1093/nar/gku1177.

Schmidt, D., P. C. Schwalie, M. D. Wilson, B. Ballester, A. Goncalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom. 2012. "Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages." *Cell* 148 (1-2):335-48. doi: 10.1016/j.cell.2011.11.058.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, I. Talianidis, P. Flicek, and D. T. Odom. 2010.

"Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding." *Science* 328 (5981):1036-40. doi: 10.1126/science.1186176.

Shibata, Y., N. C. Sheffield, O. Fedrigo, C. C. Babbitt, M. Wortham, A. K. Tewari, D. London, L. Song, B. K. Lee, V. R. Iyer, S. C. Parker, E. H. Margulies, G. A. Wray, T. S. Furey, and G. E. Crawford. 2012. "Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection." *PLoS Genet* 8 (6):e1002789. doi: 10.1371/journal.pgen.1002789.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. 2005. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." *Genome Res* 15 (8):1034-50. doi: 10.1101/gr.3715005.

Stefflova, K., D. Thybert, M. D. Wilson, I. Streeter, J. Aleksic, P. Karagianni, A. Brazma, D. J. Adams, I. Talianidis, J. C. Marioni, P. Flicek, and D. T. Odom. 2013. "Cooperativity and rapid evolution of cobound transcription factors in closely related mammals." *Cell* 154 (3):530-40. doi: 10.1016/j.cell.2013.07.007.

Sugathan, A., and D. J. Waxman. 2013. "Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver." *Mol Cell Biol* 33 (18):3594-610. doi: 10.1128/MCB.00280-13.

Sundaram, V., Y. Cheng, Z. Ma, D. Li, X. Xing, P. Edge, M. P. Snyder, and T. Wang. 2014. "Widespread contribution of transposable elements to the innovation of gene regulatory networks." *Genome Res* 24 (12):1963-76. doi: 10.1101/gr.168872.113.

Taft, R. J., M. Pheasant, and J. S. Mattick. 2007. "The relationship between non-protein-coding DNA and eukaryotic complexity." *Bioessays* 29 (3):288-99. doi: 10.1002/bies.20544.

Tamura, K., S. Subramanian, and S. Kumar. 2004. "Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks." *Mol Biol Evol* 21 (1):36-44. doi: 10.1093/molbev/msg236.

Uebbing, S., A. Kunstner, H. Makinen, and H. Ellegren. 2013. "Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers." *Genome Biol Evol* 5 (8):1555-66. doi: 10.1093/gbe/evt114.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. 2009. "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." *Genome Res* 19 (2):327-35. doi: 10.1101/gr.073585.107.

Villar, D., P. Flicek, and D. T. Odom. 2014. "Evolution of transcription factor binding in metazoans - mechanisms and functional implications." *Nat Rev Genet* 15 (4):221-33. doi: 10.1038/nrg3481.

Wiens, G. D., D. D. Rockey, Z. Wu, J. Chang, R. Levy, S. Crane, D. S. Chen, G. R. Capri, J. R. Burnett, P. S. Sudheesh, M. J. Schipma, H. Burd, A. Bhattacharyya, L. D. Rhodes, R. Kaul, and M. S. Strom. 2008. "Genome sequence of the fish pathogen Renibacterium salmoninarum suggests reductive evolution away from an environmental Arthrobacter ancestor." *J Bacteriol* 190 (21):6970-82. doi: JB.00721-08 10.1128/JB.00721-08.

Wilbanks, E. G., and M. T. Facciotti. 2010. "Evaluation of algorithm performance in ChIP-seq peak detection." *PLoS One* 5 (7):e11471. doi: 10.1371/journal.pone.0011471.

Wong, E. S., D. Thybert, B. M. Schmitt, K. Stefflova, D. T. Odom, and P. Flicek. 2015. "Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals." *Genome Res* 25 (2):167-78. doi: 10.1101/gr.177840.114.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L. H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y. C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K. H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X. Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, B. Ren, and Encode Consortium Mouse. 2014. "A comparative encyclopedia of DNA elements in the mouse genome." *Nature* 515 (7527):355-64. doi: 10.1038/nature13992.

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, A. Odeen, J. Cui, Q. Zhou, L. Xu, H. Pan, Z. Wang, L. Jin, P. Zhang, H. Hu, W. Yang, J. Hu, J. Xiao, Z. Yang, Y. Liu, Q. Xie, H. Yu, J. Lian, P. Wen, F. Zhang, H. Li, Y. Zeng, Z. Xiong, S. Liu, L. Zhou, Z. Huang, N. An, J. Wang, Q. Zheng, Y. Xiong, G. Wang, B. Wang, J. Wang, Y. Fan, R. R. da Fonseca, A. Alfaro-Nunez, M. Schubert, L. Orlando, T. Mourier, J. T. Howard, G. Ganapathy, A. Pfenning, O. Whitney, M. V. Rivas, E. Hara, J. Smith, M. Farre, J. Narayan, G. Slavov, M. N. Romanov, R. Borges, J. P. Machado, I. Khan, M. S. Springer, J. Gatesy, F. G. Hoffmann, J. C. Opazo, O. Hastad, R. H. Sawyer, H. Kim, K. W. Kim, H. J. Kim, S. Cho, N. Li, Y. Huang, M. W. Bruford, X. Zhan, A. Dixon, M. F. Bertelsen, E. Derryberry, W. Warren, R. K. Wilson, S. Li, D. A. Ray, R. E. Green, S. J. O'Brien, D. Griffin, W. E. Johnson, D. Haussler, O. A. Ryder, E.

Willerslev, G. R. Graves, P. Alstrom, J. Fjeldsa, D. P. Mindell, S. V. Edwards, E. L. Braun, C. Rahbek, D. W. Burt, P. Houde, Y. Zhang, H. Yang, J. Wang, Consortium Avian Genome, E. D. Jarvis, M. T. Gilbert, and J. Wang. 2014. "Comparative genomics reveals insights into avian genome evolution and adaptation." *Science* 346 (6215):1311-20. doi: 10.1126/science.1251385.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol* 9 (9):R137. doi: 10.1186/gb-2008-9-9-r137.

Zhu, L. J., R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe, and M. H. Brodsky. 2011. "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." *Nucleic Acids Res* 39 (Database issue):D111-7. doi: 10.1093/nar/gkq858.

CHAPTER 2: Epigenetic aging signatures in mice are slowed by dwarfism, calorie restriction and rapamycin treatment

2.1 Abstract

Background. Global but predictable changes impact the DNA methylome as we age, acting as a type of molecular clock. This clock can be prematurely advanced by conditions that decrease lifespan, raising the question of whether it can also be slowed, for example by conditions that increase lifespan. Mice are particularly appealing organisms for studies of mammalian lifespan; however, epigenetic clocks have thus far been formulated only in humans.

Results. We first examine whether mice and humans experience similar patterns of change in the methylome with age. We find moderate conservation of CpG sites whose methylation is altered with age, with both species showing an increase in methylome disorder during aging. Based on this analysis, we formulate an epigenetic-aging model in mice using the liver methylomes of 107 mice from 0.2 to 26.0 months old. To examine whether epigenetic aging signatures are slowed by longevity-promoting interventions, we analyze 28 additional methylomes from mice subjected to lifespan-extending conditions, including Prop1$^{df/df}$ dwarfism, calorie restriction or dietary rapamycin. We find that mice treated with these lifespan-extending interventions are significantly younger in epigenetic age than their untreated, wild type age-matched controls. Thus, age-related methylation changes are mitigated by dwarfism, calorie restriction and rapamycin treatment.

Conclusions. This study shows that lifespan-extending conditions can slow molecular changes associated with an epigenetic clock.

2.2 Introduction

In humans, numerous CpG sites have DNA methylation states that correlate with age. These associations have been used to formulate models, called epigenetic clocks, that make quantitative predictions of age based on selected sets of CpG sites (Hannum et al. 2013; S. Horvath 2013; Weidner et al. 2014). These models are derived from the methylation profile of many individuals measured using oligonucleotide arrays, such as the Illumina 450K platform, which determines the levels of methylation value at >450,000 CpG sites genome-wide. Although the age predictions of these molecular models are generally very accurate across the human population, for particular individuals the prediction can be markedly different from the actual chronological age. For example, an advanced molecular age relative to chronological age has been associated with a number of diseases, such as obesity, viral infection and Down syndrome (S. Horvath et al. 2014; Gross et al. 2016; Steve Horvath et al. 2015). Furthermore, a recent retrospective analysis of longitudinal cohort studies has shown that a molecular age advancement of 5 years corresponded to a 21% increased risk of mortality overall (Marioni et al. 2015). Thus, predictions of "epigenetic age" may be an indication of an individual's biological state of aging.

Beyond these examples of advanced epigenetic aging, a complementary but unanswered question is whether epigenetic clocks can also be slowed. Epigenetic aging studies in humans have not been well-suited to address questions of slowed aging, given the lack of well-documented interventions that enhance health or lifespan and the difficulty of controlling for confounding factors. On the other hand, rodents are particularly appealing experimental organisms in studies of mammalian aging, since they are genetically tractable and can be subjected to potential lifespan-extending interventions. The earliest described such intervention, calorie restriction, was shown to extend rodent lifespan by as much as two-fold (McCay et al. 1935). These findings have since been replicated in numerous mouse strains (Means, Higgins, and Fernandez 1993). Another well-studied lifespan-extending condition is a single-point mutation in the Prop1 gene that results in dwarfism and lifespan extension up to 1.5-fold (Brown-Borg et al. 1996). These effects are likely due to reduced somatotropic signaling (Bartke and Brown-Borg 2004). A more recently described treatment, dietary rapamycin, has been reported to increase lifespan of genetically heterogeneous mice by 1.2-fold (Miller et al. 2014).

Despite these known lifespan-extending interventions, an epigenetic clock has not yet been formulated for mice. Nonetheless, mouse methylation signatures can now be analyzed genome-wide using either Reduced Representation Bisulfite Sequencing (RRBS) or Whole Genome Bisulfite Sequencing (WGBS) (Laird 2010). Using such data, previous studies have suggested that mice might experience patterns of epigenetic aging similar to those documented in humans (Avrahami et al. 2015; Sun et al. 2014; Spiers et al. 2016). For instance, CpG methylation sites distinguish young versus old mouse hematopoietic stem cells (Beerman et al. 2013), and CpG methylations altered in murine acute myeloid leukemia are also found to change with age (Maegawa et al. 2014). These findings suggest that an epigenetic measure of age is plausible for mice.

Here, we ask if conditions that extend mouse lifespan – Prop1$^{df/df}$ dwarfism, calorie restriction and dietary rapamycin – also affect a mouse epigenetic clock. It is possible that these lifespan-extending conditions operate independently of the changes that underlie an epigenetic clock, which would then proceed at a normal rate despite these interventions. To distinguish between these possibilities, we first assess whether there are similarities between mice and human epigenetic aging. We then formulate epigenetic readouts of age that are used to score the effect of lifespan-extending interventions.

2.3 Results

Age-related methylation changes share common behavior in mouse and human**.** First, we assessed the similarities of age-related methylome changes between mice and humans. For this purpose, we obtained publicly available mouse methylation data from Reizel *et al.* (Reizel et al. 2015) consisting of RRBS from livers of 102 male or female C57BL/6 mice ranging in age from 0.2 to 7.1 months. These data were filtered to identify sites that were reliably measured with sufficient sequencing depth in most mice (**Methods**), yielding 36,094 CpG sites total.

To compare these sites to human, we found that 27,612 CpG sites were conserved in humans. Next, we obtained publicly available methylation data from 164 human livers that were generated using 450K Illumina methylation arrays (Ahrens et al. 2013; S. Horvath et al. 2014). We identified 2,634 CpG sites that were assayed in both the human Illumina arrays and were orthologous from the set of filtered sites from mouse RRBS (**Figure 2.1a**). From this orthologous-profiled space, we identified 88 age-associated sites in mice (for which the methylation status had a significant association with age) and 176 age-associated sites in humans (**Figure 2.1a**, likelihood ratio test at 1% FDR, see **Methods**). Among these, we saw slight but significant overlap between sites that were age-associated in mice versus sites that were age-associated in humans (**Figure 2.1a**) ($p < 0.01$ by hypergeometric test). Notably, the age-associated sites in both species showed similar under/over enrichments in various genomic annotations, including regions marked by histones (H3K27me3, bivalent and H3K9ac). However, different genomic regions were associated with statistical significance in mice and humans, with only H3K27ac regions significantly under-enriched in both species (**Supplementary Figure S2.1**). Thus, it seems that age-associated CpG sites in mice and humans are moderately conserved with respect to various genomic regions affected.

Previous methylation studies of whole blood in humans have observed increasing entropy with age (Hannum et al. 2013; Gross et al. 2016). Increasing entropy indicates that, during aging, the state of each CpG becomes less uniform across the cell population [1]. We asked if this trend of increasing disorder of age-associated CpG sites in the methylome also existed in mice and human livers, regardless if a particular site was sampled in both species. We saw that, in both mice and human, the age-associated regions of the methylome tended towards higher disorder (**Figures 2.1b**,**c**). This suggests that a trend towards disorder over time is a conserved property of aging in mammals.

<u>Development of an epigenetic clock in mice.</u> Motivated by the shared patterns affecting the aging epigenome in mice and humans, we next formulated an epigenetic clock for mice. Towards this goal, we created a consolidated mouse liver methylome dataset combining two previous studies (Reizel et al. 2015; Gravina et al. 2016) with data newly generated in this study (**Supplementary File S2.1**). This consolidated dataset consisted of 107 liver methylomes of mice aged 0.2 to 26.0 months old (**Supplementary Figure S2.2a**), covering a total of 7,628 CpG sites that were detected in nearly all samples (**Methods**). Normalization with ComBat (Leek et al. 2012; Johnson, Li, and Rabinovic 2007) was used to estimate and remove effects resulting from the different sequencing technologies (RRBS and WGBS) and mouse strains (Ames, C57BL/6 and UM-HET3) used in this integrated dataset (**Methods**). To train a predictive model of mouse age, which can be used as an epigenetic clock, we then applied ElasticNet (Zou and Hastie 2005), a statistical regression framework used previously to formulate epigenetic clocks in humans (Hannum et al. 2013; S. Horvath 2013). This training process selected a subset of 148 CpG sites for an epigenetic clock in mice livers (**Supplementary File S2.2**).

We pursued two different strategies to assess predictive performance. First, we performed four-fold cross validation, in which the 107 mice used for training were arbitrarily divided into four sets of comparable sizes. Each of these sets was withheld, in turn, from model training and instead used to test the performance of the trained model. In this cross-validation scenario, we found that the ages of the test sets were accurately predicted with a correlation ranging from 83% to 92% (average r = 0.91; **Figure 2.2a**). Second, we tested the performance of the model when predicting age from the liver methylomes of 50 mice that had not been used for model training or cross validation (spanning three mouse strains and two ages, 2 and 22 months, **Supplementary File S2.1: Datasets used summary**). Predicted epigenetic ages were well correlated with chronological ages (**Figure 2.2b**) and did not show any strain-specific effects: 2-month-old Ames wild type, UM-HET3 and C57BL/6 mice had roughly the same epigenetic age; the same was true for 22-month-old Ames wild type and untreated UM-HET3 mice (**Supplementary Figure S2.2b**, **Supplementary File S2.3: Wild type mice predictions & stats**).

Lifespan extension slows epigenetic aging. We next examined in greater detail the behavior of the 148 CpG sites used as features by principal component analysis (**Supplementary Figure S2.3A**). The first principal component of these features (PC1) correlated strongly with age, and PC1 values of mice subjected to lifespan-extending treatments were always less than PC1 values of age-matched controls (**Figures 2.3a,b**, **Supplementary Figure S2.3a**, **Supplementary Table S2.1**).

We next applied the epigenetic-aging model to WGBS data generated for mice subjected to various lifespan-extending conditions (**Methods, Supplementary File S2.1: Datasets used summary & Supplementary File S2.3: Long-lived mice predictions**). This analysis included methylomes from Prop1$^{df/df}$ dwarf mice aged 2 or 22 months (Brown-Borg et al. 1996), with four in each group, four calorie-restricted mice aged to 22 months, and four rapamycin-treated mice aged to 22 months (Miller et al. 2014). We found that the predicted epigenetic ages of these long-lived mice were significantly less than those of age-matched control mice (**Figure 2.3c**). Reinforcing this observation, such differences were also detected by an ANOVA statistical analysis between the lifespan-extending conditions versus control mice aged to 22 months ($p < 10^{-4}$, **Methods, Supplementary File 2.3: Treatment vs wild type stats**). In particular, an average reduction of 10.1 months was seen when comparing the epigenetic ages of 22-month-old dwarf mice to 22-month-old wild types ($p < 0.01$ by t-test, **Figure 2.3d**). Similar reductions in epigenetic ages were observed in calorie-restricted mice versus their age-matched controls, corresponding to a 9.4-month decrease on average ($p < 10^{-4}$, **Figure 2.3d**). Rapamycin-treated mice had a smaller, but significant effect on epigenetic ages, which corresponded to a 6.0-month decrease on average compared to their age-matched controls ($p < 0.05$, **Figure 2.3d**). Finally, 2-month-old dwarf mice also had reduced epigenetic ages compared to 2-month-old wild type mice, by 1.5 months on average ($p < 10^{-3}$, **Figure 2.3d**). These results are consistent with the smaller magnitudes of age-associated PC1 of long-lived mice, relative to their age-matched controls.

We next assessed the change in methylation over age of the 148 CpG sites used to formulate this epigenetic clock. Among these CpG sites, we found that 76 gained methylation over age and 72 lost methylation over age. These sites clustered the mice mostly according to age and treatment rather than by genetic background (**Figure 2.3e**, **Supplementary Figure S2.3b**). Among CpG sites whose methylation decreased with age, we saw that long-lived mice generally had higher methylation values than the age-matched controls, which may contribute towards the observed decrease in epigenetic ages (**Figure 2.3e**). Thus, whether examined individually (**Figure 2.3e**) or summarized along a single dimension (**Figure 2.3a,b**), changes in methylation due to aging are generally less extreme in mice exposed to pro-longevity conditions, leading to younger epigenetic ages relative to their age-matched controls (**Figure 2.3c,d**).

2.4 Discussion

Previous studies in human have shown that epigenetic clocks can be accelerated by conditions associated with decreased lifespan (Gross et al. 2016; S. Horvath et al. 2014; Steve Horvath et al. 2015). However, it was unclear if these epigenetic clocks can be slowed by conditions that increase lifespan. Here, we have found that lifespan-extending interventions can indeed slow an epigenetic clock in mice livers. Previous studies of these longevity-promoting interventions have shown that these interventions not only extend lifespan (Brown-Borg et al. 1996; Means, Higgins, and Fernandez 1993; Harrison et al. 2009), but also improve tissue and physical functioning over age (Bartke and Westbrook 2012; Arum et al. 2013; Masternak et al. 2009; Wilkinson et al. 2012). Interestingly, rapamycin had a smaller effect than the other treatments considered here, possibly due to metabolic differences, such as increased insulin resistance under rapamycin treatment (Lamming et al. 2012). Nonetheless, our findings suggest that epigenetic clocks, measured from DNA methylation, can be slowed by lifespan-extending conditions.

Notably, we found that dwarf mice had a decreased epigenetic age at our earliest time-point, when just 2-months-old. (**Figure 2.3c,d**). This finding suggests that age-related changes in the methylome occur during both development and aging (Zampieri et al. 2015; Johansson, Enroth, and Gyllensten 2013; Takasugi 2011; Medvedev 1990). Prominent changes in the DNA methylome begin during development in mice and continue throughout adulthood (Takasugi 2011). In humans, epigenetic clocks are accurate in both adolescents and adults (S. Horvath 2013).

Thus, the decrease in epigenetic age of young dwarf mice is consistent with the apparent developmental delay observed in dwarf mice (Bartke and Brown-Borg 2004).

When comparing age-related methylation changes between mice and humans, we found moderate conservation for age-associated CpG sites and genomic regions. Most strikingly, we found that the age-associated methylome exhibits increased disorder in both species (**Figure 2.1b,c**). These results suggest that, regardless of the specific regions impacted, the increased disorder of the age-associated methylome is a common feature of mammalian aging. This increased disorder of the age-associated methylome may explain how the methylome is strongly associated with chronological aging (Hannum et al. 2013; S. Horvath 2013; Weidner et al. 2014; Lopez-Otin et al. 2013).

Finally, since this mouse clock was developed using liver methylomes, in future studies it will be very interesting to examine whether these clocks are similar across various tissues. Intriguingly, previous studies in humans have found that obesity is specifically associated with epigenetic age advancement in the liver but not in other tissues such as blood (S. Horvath et al. 2014). Furthermore, rapamycin treatment has been shown to accelerate cataract formation in eyes and increase testicular degeneration, but delays age-related phenotypes in other tissues (Wilkinson et al. 2012). A key question will be whether these same tissue-specific effects are reflected in epigenetic aging rates, where some tissues may reflect slowed aging while others reflect accelerated aging.

2.5 Conclusions

In summary, we have formulated an epigenetic-aging model in mice and used it to find evidence that lifespan-extending conditions slow an epigenetic clock in mice livers. To further understand whether lifespan-extending conditions promote more youthful epigenetic signatures globally, it will be of interest to study different tissues, as well as profile mice exposed to other lifespan-extending conditions, such as methionine restriction or other mutations in somatotropic signaling pathways (Bartke and Westbrook 2012). Ultimately, such studies will help elucidate the relationship between the slowed epigenetic clock and healthy aging.

2.6 Methods

Long-lived mice. To study the effects of dwarfism, we studied 2 or 22-month-old male Ames Prop1$^{df/df}$ dwarf and wild-type mice livers (Brown-Borg et al. 1996), with four mice in each group. Mice were maintained under controlled conditions at the University of North Dakota with access to food *ad libitium*. To study the effects of calorie restriction and rapamycin treatment, we used female UM-HET3 mice livers aged to 22 months, where mice were either subjected to calorie restriction (60% of food consumption relative to age-matched controls, gradually reduced over 2 weeks) or 42 mg/kg dietary rapamycin treatment from 4 – 22 months, or left untreated, with four mice in each group. We also obtained livers from female untreated UM-HET3 mice livers aged to 2 months (Miller et al. 2014). UM-HET3 mice were maintained at the University of Michigan. Weights of these mice are described in **Supplementary Table S2.2**.

WGBS library preparation. DNA was isolated from mice livers using the DNeasy blood and tissue kit (Qiagen). Whole-genome bisulfite sequencing was carried out by the Beijing Genomics Institute (Shenzhen) following standard protocols (Urich et al. 2015). Briefly, DNA was fragmented using sonication to an average fragment size of 100-300bp, end-repaired, and ligated to methylated-sequencing adapters to generate sequencing libraries. Bisulfite conversion was performed on these sequencing libraries using the ZYMO EZ DNA Methylation-Gold kit and sequenced using 90bp paired-end sequencing on an Illumina HiSeq-4000. Ames mice were sequenced to an expected 15x coverage; UM-HET3 mice were sequenced to an expected 5x coverage.

Data processing. For the WGBS study in long-lived mice, sequencing reads were trimmed using Trim Galore ("Babraham Bioinformatics - Trim Galore!" 2016), aligned to a bisulfite-converted mouse genome (mm9) obtained from UCSC (Hinrichs et al. 2006) using bowtie (Langmead et al. 2009), and methylation states were called using bismark v0.10.0 (Krueger and Andrews 2011). The resulting sites were then converted to mm10 coordinates using liftOver (Hinrichs et al. 2006) with default parameters.

In addition to the above data, public bisulfite sequencing data were downloaded from GEO (Barrett et al. 2013) or the Sequence Read Archive (SRA) using the following accession numbers: GSE60012 (Reizel et al. 2015), GSE52266 (Cannon et al. 2014), GSE67507 (Orozco et al. 2015) and SRA344045 (Gravina et al. 2016). Sequencing reads were trimmed using Trim Galore ("Babraham Bioinformatics - Trim Galore!" 2016) with default parameters, aligned to bisulfite-converted Ensembl mmGRC38 version 84 (Yates et al. 2016) using bowtie2 (Langmead and Salzberg 2012) with parameters -N 1, and the methylation states were determined using Bismark v0.14.3 (Krueger and Andrews 2011). When multiple sequencing runs were associated with a single sample, the methylation states for each CpG were collapsed by summing the reads.

Human 450K liver data were downloaded from GEO using accession numbers GSE61258 and GSE48325 corresponding to Horvath *et al.* (S. Horvath et al. 2014) and Ahrens *et al.* (Ahrens et al. 2013) datasets. The data were processed in R using Minfi (Aryee et al. 2014). Missing data were imputed using impute package in R (Hastie et al. 2011). The data were then beta-mixture quantile normalized (Teschendorff et al. 2013) using a gold reference distribution following the procedure provided by Horvath (S. Horvath 2013). The gold reference distribution was set to the mean probe values from GSE61258.

Evolutionary trends. For the comparison of mice to humans, we wanted to maximize the number of mouse CpG markers that we could compare reliably across species. For this reason, we limited our analysis to RRBS datasets obtained from GEO. Specifically, we filtered Reizel *et al.* (Reizel et al. 2015) with Cannon *et al.* (Cannon et al. 2014) and Orozco *et al.* (Orozco et al. 2015) to identify reproducible CpG sites. Sites were filtered according to the following criteria: ≥ 5 reads, < 20% missing data across mice from all three studies, and distinct mapping onto chromosomes 1-19. We then removed individual mouse samples missing > 40% of these sites. These filtering steps resulted in 97 samples profiled across 36,094 sites in Reizel *et al.* (Reizel et al. 2015). Missing data were imputed using the mean methylation value for that site.

To define a commonly profiled set of orthologous CpG sites, we mapped the 36,094 sites profiled in mm10 to hg19 coordinates using liftOver (Hinrichs et al. 2006), with -minMatch = 0.1. The resulting coordinates were intersected with the Illumina 450K probes, as defined by their locations from the Illumina manifest (bedtools intersectbed (Quinlan and Hall 2010)). Any mouse sites that mapped to the same human site were combined by taking the average value of these sites.

Annotation tracks were downloaded from Encode for human hepatocytes from UCSC (Rosenbloom et al. 2013). The following data tracks were downloaded: DNASE-seq, H3K36me3, H3K4me1, H3K27ac, H3K9ac, H3K4me3 and H3K27me3. Enhancer regions were defined as the intersected regions between H3K27ac and H3K4me1. Bivalent regions were defined as the intersected regions between H3K4me3 and H3K27me3. Repeat elements were downloaded from UCSC for hg19 (Rosenbloom et al. 2015). CpG sites were mapped to each feature by intersecting the site coordinates with each annotation using bedtools intersectbed. Annotations for TSS, 5'UTR, body, exons, shelf, island and shore were defined by the Illumina 450K manifest. Promoters were defined as CpG sites with TSS annotations. Similarly for mice, annotation tracks were downloaded from UCSC for the same marks from adult male mice liver. Gene features for mice were also downloaded from UCSC for mm10 (Rosenbloom et al. 2015) and assigned to sites using bedtools intersectbed. Promoters in mice were defined as 2kb upstream of protein-coding transcripts. We only considered annotations that fell within the orthologous profiled set of CpGs. These annotations are used as genomic regions in the following section.

Odds ratios were calculated by counting orthologous CpGs sites that fell into the following categories: age-genomic region, not age-genomic region, age-not genomic region and not age-not genomic region. This process was repeated for each genomic region separately in both human and mouse. When there were overlapping genomic region annotations for sites, sites were counted only for the genomic region considered so that sites are not counted twice. Over-represented genomic regions are those with an OR > 1 and under-represented genomic regions are those with an OR < 1. P-values were calculated using Fisher's exact test.

To identify age-associated sites, we built a multivariate linear model regressing each methylation site against treatment, gender and age in mice, or against BMI, gender and age in humans. Then, we conducted a drop-one F-test to determine if age had a significant association with that site. For comparisons in the orthologous profiled space between mice and humans, we conducted the drop-one F-test using Reizel *et al.* (Reizel et al. 2015) for mice or all human samples, and we selected sites that had an age-association at a Benjamini-Hochberg 1% FDR. To calculate the significance of the overlap, we used a hypergeometric test.

To identify all age-associated sites, regardless of conservation, we conducted the same drop-one F-test, first using the 97 mice of Reizel *et al.* (Reizel et al. 2015) for all 36,094 CpG sites, then selecting CpG sites that passed a Benjamini-Hochberg 1% FDR. We repeated this analysis using the 2.1-month mice from Cannon *et al.* (Cannon et al. 2014) and 3.7-month mice from Orozco *et al.* (Orozco et al. 2015), using the CpG sites identified in Reizel *et al.* (Reizel et al. 2015), and selected sites that continued to have an age-association at a Benjamini-Hochberg 1% FDR. Using this criteria, we found 393 age-associated sites in mice. These sites were used to calculate entropy for Reizel *et al.* (Reizel et al. 2015) (**Figure 2.1b**). We identified age-associated CpG sites in humans similarly, using all 485,512 CpG sites on the 450K Illumina chip, first in GSE61258 (S. Horvath et al. 2014) (79 samples), identifying CpG sites with an age-association at a Benjamini-Hochberg 1% FDR threshold. We repeated this analysis for the identified CpG sites in GSE48325 (Ahrens et al. 2013) (85 samples), selecting CpG sites that passed a Benjamini-Hochberg 1% FDR threshold. Using these criteria, we found 322 age-associated CpG sites. These sites were used to calculate entropy (**Figure 2.1c**) for GSE61258 (S. Horvath et al. 2014).

Entropy was calculated according to the formula described by (Hannum et al. 2013), calculated with the following formula:

$$Entropy = \frac{1}{N * log(\frac{1}{2})} \sum_{i}^{N} [MF_i * \log(MF_i) + (1 - MF_i) * \log{(1 - MF_i)}]$$

where MF_i is the methylation fraction of the i^{th} methylation CpG site and N is the number of age-associated CpG sites (393 sites for mice and 322 sites for human, described above). Since the value of entropy approaches 0 when MF_i approaches 0, the methylation sites with a value of 0 were set to 0.

Epigenetic clock data processing and data normalization. For construction of an epigenetic-aging model, which we use as a mouse epigenetic clock, we used GSE60012 (Reizel et al. 2015), SRA344045 (Gravina et al. 2016) and our own control mice to identify sites that were profiled across all studies, for a total of 124 mice liver/hepatocyte samples. These studies represent mice profiled across multiple time-points. Since RRBS is targeted towards CpG rich regions of the genome, we included sites that were covered by ≥ 2 reads in 97% of mice, mapped to chromosomes 1-19 and had a standard deviation > 0 and ≤ 20%. Mice missing over 30% of these sites were removed from further analysis. Missing data were imputed by the mean value of each site. These filtering steps resulted in 119 samples profiled across 7,628 CpG sites. For studies profiling a single time point (Cannon et al. 2014) and the long-lived mice, in order to maximize the overlap with the 7,628 CpG sites selected above, we considered any site with ≥ 1 reads (bedtools intersectbed). Missing data were imputed identically as described above.

All data were then normalized using ComBat (nonparametric mode) from the SVA package in R (Leek et al. 2012; Johnson, Li, and Rabinovic 2007). Ages (in days) were transformed to $\log_2$ scale, prior to normalization. The specific sequencing studies ((Reizel et al. 2015; Gravina et al. 2016; Cannon et al. 2014), Ames and UM-HET3) were used to represent batch, and the model provided to ComBat included the covariates age, gender and treatment. After performing ComBat, we used principal component analysis to verify that this normalization reduced the effects due to differences in sequencing technology or mouse strains (**Supplementary Figure S2.2c,d**). Bismark alignment reports, as well as average read depth per unique CpG called and per CpG used to construct the epigenetic-aging model, are shown in **Supplementary File S2.1: Public data and Data here detailed**.

<u>Epigenetic-aging model construction.</u> The normalized methylation values of (Reizel et al. 2015; Gravina et al. 2016; Cannon et al. 2014) and data from wild type, untreated UM-HET3 and Ames aged to 2 and 22 months (one from each group) (**Supplementary File S2.1: Datasets used summary**), were used as training data for ElasticNet regression (Zou and Hastie 2005) using the python scikit-learn package (Pedregosa et al. 2011). The normalized methylation values were used

as features, and the $\log_2$ transformed ages (in days) was used as the predicted variable. Model fitting parameters were selected using four-fold cross validation. The final model was trained on these training data with the most optimal regularization parameters when averaging the four-fold cross validation results. The model sites selected by ElasticNet, along with the associated weights and intercept, are shown in **Supplementary File S2.2**.

We assessed whether epigenetic ages were informative by comparing the epigenetic ages for untreated, wild type mice from our study or mice from *Cannon et al.* (Cannon et al. 2014). We used either a t-test or an ANOVA to compare whether epigenetic ages were significantly different between 2 versus 22-month-old mice, and whether epigenetic ages of mice with similar chronological ages were affected by differences in genetic backgrounds (**Supplementary File S2.3: Wild type stats)**. To assess the effect of normalization in addition to selection of regularization parameters or hidden biases correlated to aging signals, the covariates of each study were shuffled, ComBat normalization was repeated, and models were learned using the same strategy described above. This process was repeated 120 times and predictions between models generated from permuted data or actual data were compared using the residual (epigenetic age - chronological age) for wild type mice. The model learned from actual data minimized the residual for the wild type mice (**Supplementary Figure S2.2e-i**).

Assessing epigenetic age in long-lived mice. The epigenetic-aging model was applied to the methylation profiles of long-lived mice and the age-matched controls not used for training (**Supplementary File S2.1: Datasets used summary**). Reductions in age were calculated by subtracting the epigenetic ages of the untreated, wild type mice from those of the long-lived mice of the same genetic background. To assess the significance, we used an ANOVA for all 22-month-old mice or only 22-month-old UM-HET3 mice. We also compared the epigenetic ages between treatments with their age-matched control from the same genetic background using a t-test (**Supplementary File S2.3: Treatment vs wild type stats)**.

Principal component analysis (PCA). PCA was conducted using the scikit-learn package from the 148 CpG sites used in the epigenetic clock. The first two PCs separated age and treatment (**Supplementary Figure S2.3A**). We assessed significance of variables that contributed to the variance along principal component 1 for each genetic background using a multivariate linear regression according to the following model:

$$Principal\ Component\ 1 \sim Age + Treatment$$

where $Treatment$ was modeled as a categorical variable and results are shown in **Supplementary Table 2.1**.

Hierarchical clustering. Hierarchical clustering was performed using python SciPy with linkage method 'average' and Euclidean distance (Jones et al. 2015). Methylation values were transformed using standard_scale=True and visualized using seaborn (Waskom et al. 2014). Hierarchical clustering was performed either using the top 20 most variable CpG sites (determined from the long-lived mice and wild type mice) or all sites used by the epigenetic-aging model.

Availability of data and material. Accession numbers for publicly available mice bisulfite datasets: GSE52266, GSE60012 GSE67507 and SRA344045; Accession numbers for human methylome datasets: GSE61258 and GSE48325; Accession numbers for long-lived and control mice bisulfite datasets: GSE89275; Scripts for analysis can be found at https://github.com/twangsd/Epigenetic_Aging_Mouse_livers

2.7 Figures

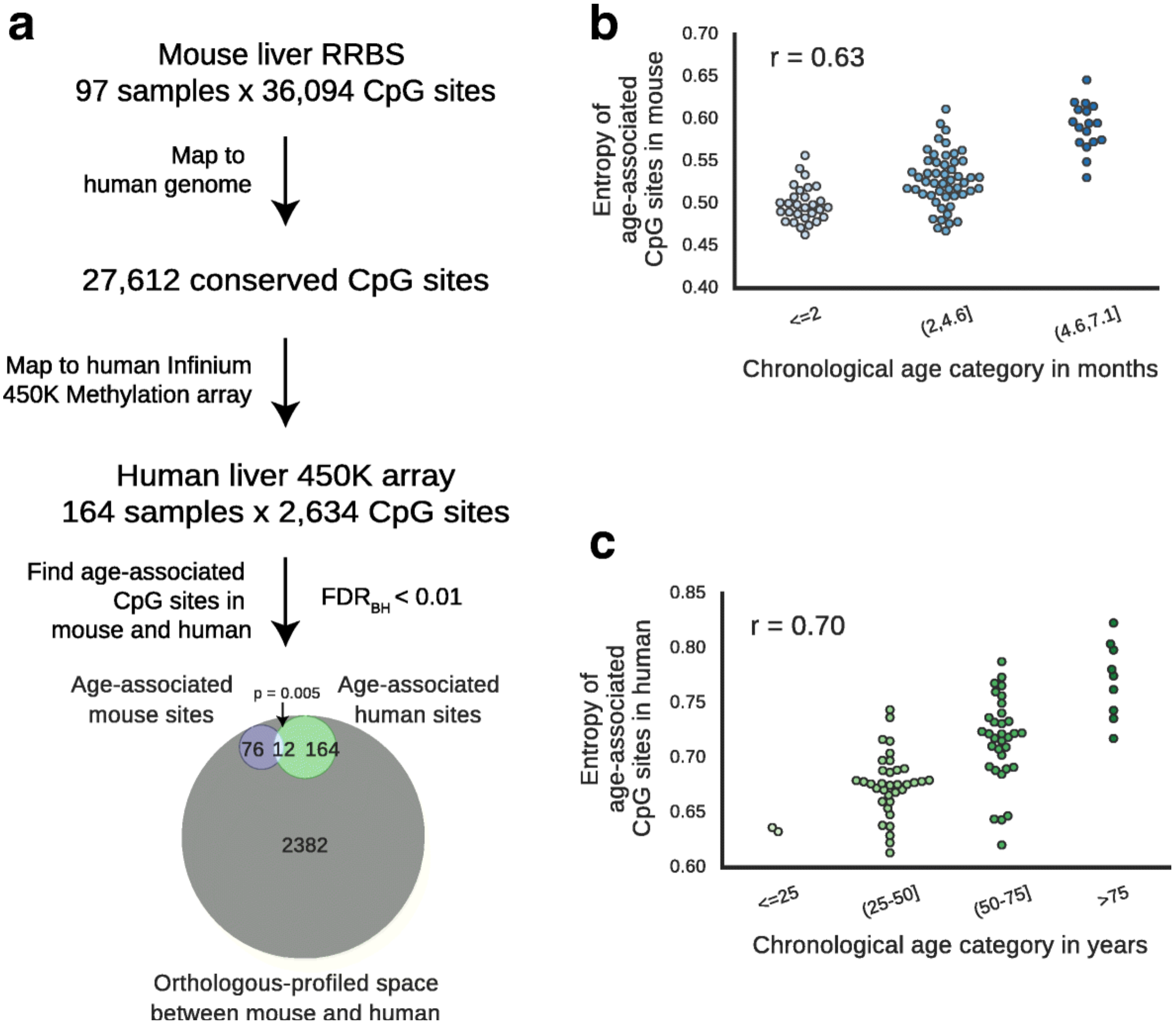


Figure 2.1: Comparison of methylation aging in mice and human livers.
a) Mapping from mouse CpG sites profiled by RRBS to orthologous CpG sites profiled by Illumina 450K human methylation array. Detailed procedures can be found in **Methods.** The Venn diagram describes the age-associated sites in the orthologous profiled space. **b-c**) Entropy across all age-associated sites in mouse (**b**) and in humans (**c**) are plotted over age. Pearson's correlation (r) is displayed (mouse $p < 10^{-11}$, human $p < 10^{-11}$).

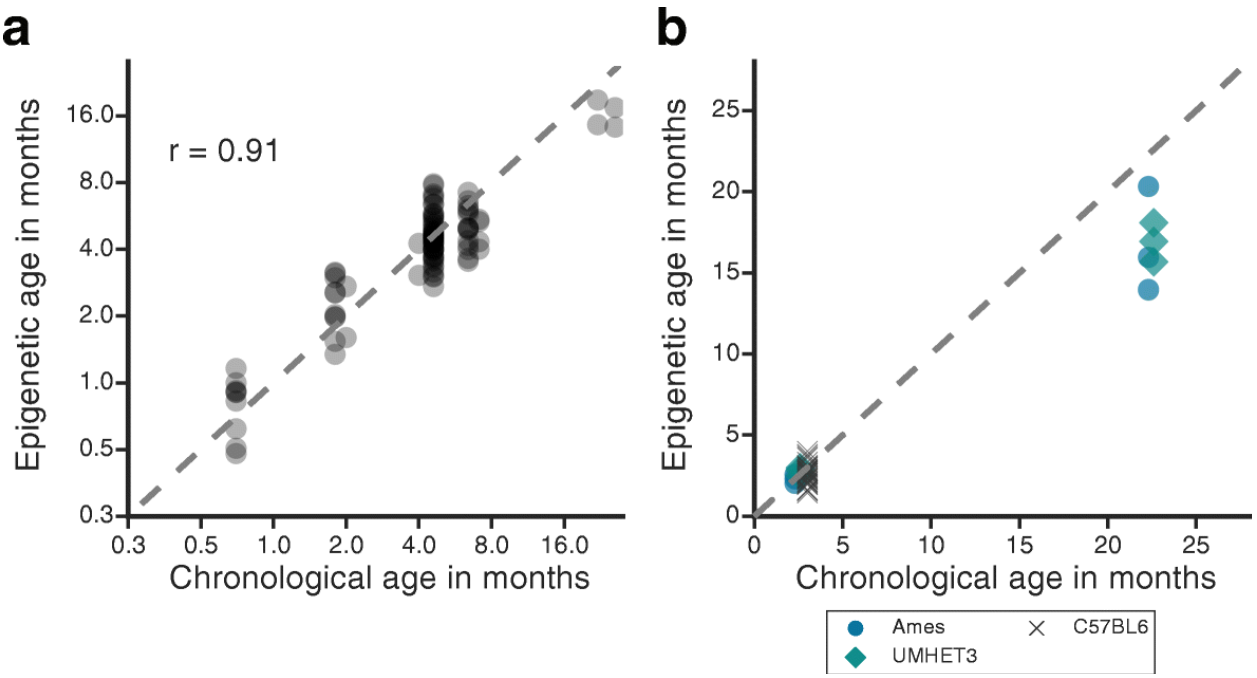
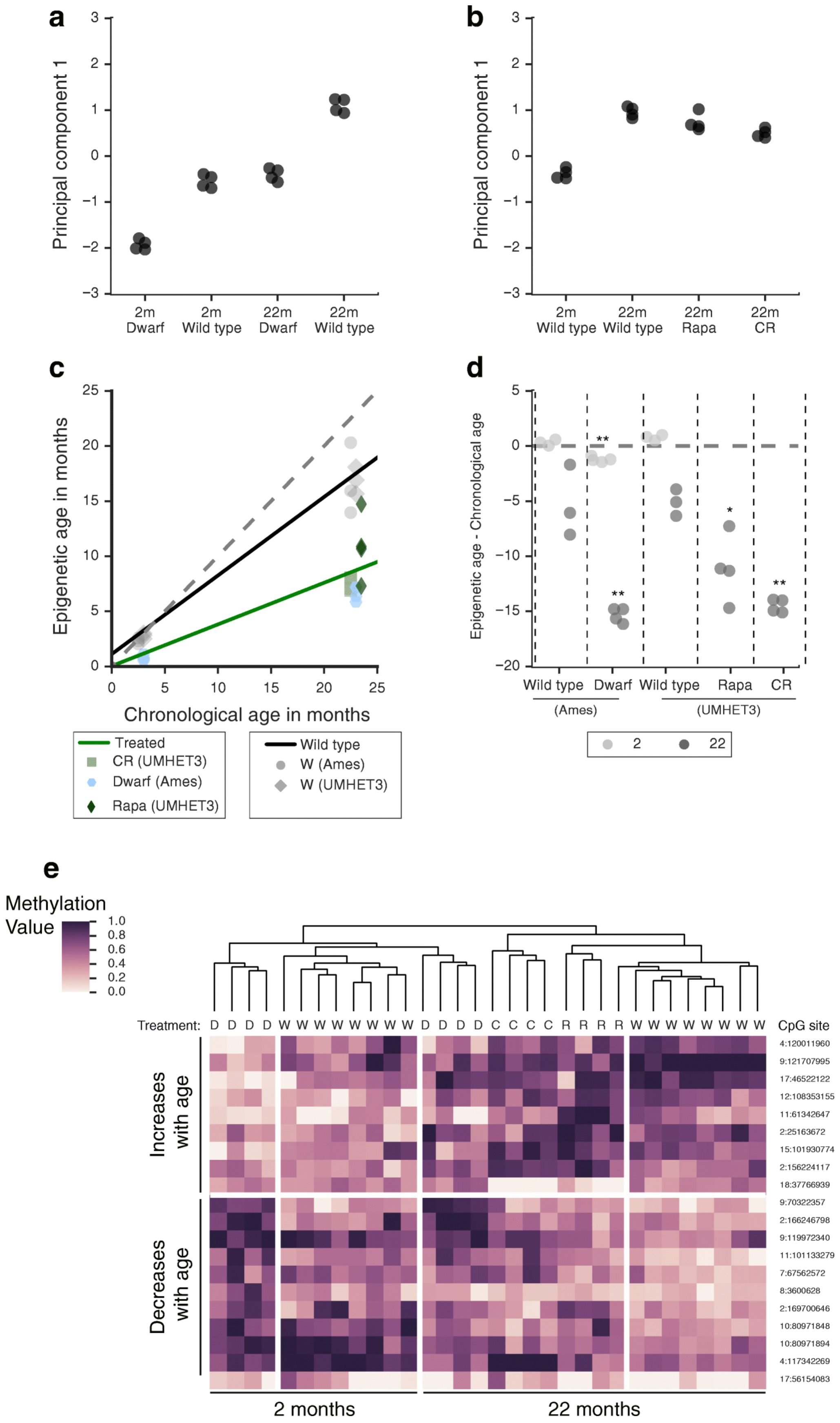


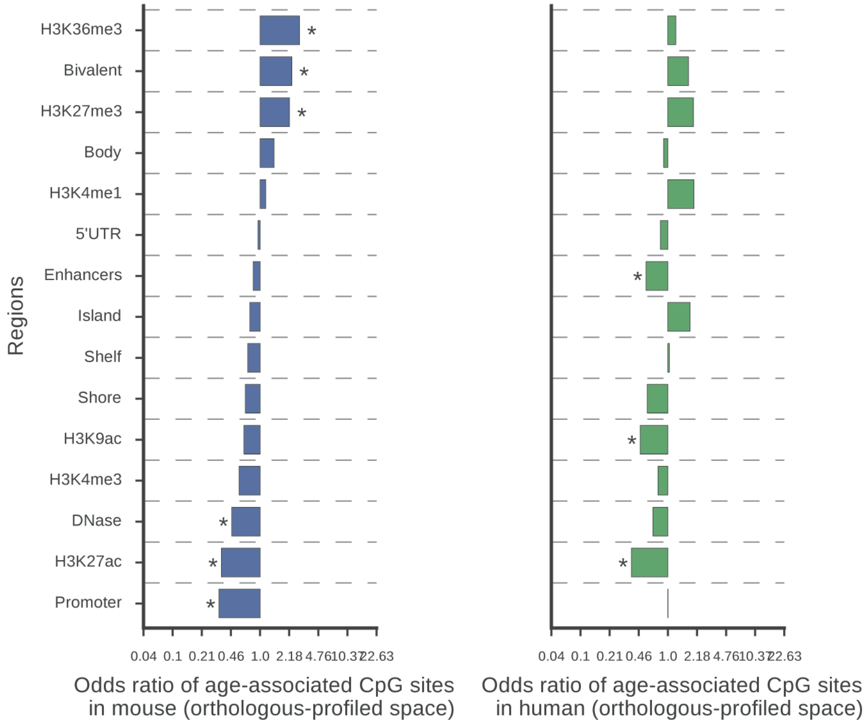
Figure 2.2: Validation of an epigenetic-aging model in mice livers.
a) Four-fold cross validation of the age predictions (y-axis, "Epigenetic age") versus chronological age (x-axis) in $\log_2$ scale. Each dot represents a prediction made from a fold of cross validation. Overall, each fold has a high Pearson's correlation (r) between epigenetic and chronological age, and the average amongst all folds is depicted. **b)** Epigenetic ages versus chronological ages for 50 wild type mice. Different symbols/colors are used to indicate mouse genetic background. The dashed line represents the diagonal in both plots.

Figure 2.3: Effects of lifespan extension on a mouse epigenetic clock.

a-b) The 148 CpG sites used the mouse epigenetic-aging model (used for a mouse epigenetic clock) were subjected to principal component analysis. Principal component 1 is plotted for wild type mice according to age and lifespan extension status, for wild type Ames or dwarf mice (**a**) or wild type UM-HET3, rapamycin treated, or calorie restricted (**b**). **c)** The mouse epigenetic-aging model applied to long-lived mice, where colors and shapes represent the different lifespan-enhancing conditions. The gray markers are the wild type mice (identical to **Figure 2B**), and the black line represents the linear fit of the epigenetic age versus chronological age of the wild type mice. The green line represents the linear fit of the epigenetic age versus chronological age for long-lived mice. The gray dashed line represents the diagonal **d)** The residual (epigenetic age - chronological age) is plotted for all mice according to their strain and treatment, and colors represent 2 or 22-months. P-values calculated by comparing ages of long-lived mice to age-matched controls of same genetic background using a t-test. ** indicates $p < 0.01$ and * indicates $p < 0.05$. **e)** Hierarchical clustering of the top 20 most variable sites used by this epigenetic clock using average linkage with Euclidean distance. Treatment is depicted under the dendrogram, CpG sites are to the right of the heatmap (chromosome:start, 0-based) and rows are blocked according to clusters of sites that increase or decrease methylation with age. m: months; R: Rapamycin treatment; C, CR: Calorie restriction, D: Ames Dwarf, W: wild type Ames or untreated, wild type UM-HET3.


a
Principal component 1
3
2
1
0
−1
−2
−3
2m Dwarf
2m Wild type
22m Dwarf
22m Wild type
b
Principal component 1
3
2
1
0
−1
−2
−3
2m Wild type
22m Wild type
22m Rapa
22m CR
c
Epigenetic age in months
25
20
15
10
5
0
0
5
10
15
20
25
Chronological age in months
Treated
CR (UMHET3)
Dwarf (Ames)
Rapa (UMHET3)
Wild type
W (Ames)
W (UMHET3)
d
Epigenetic age - Chronological age
5
0
−5
−10
−15
−20
**
**
*
**
Wild type
Dwarf
Wild type
Rapa
CR
(Ames)
(UMHET3)
2
22
e
Methylation Value
1.0
0.8
0.6
0.4
0.2
0.0
Treatment:
D D D D W W W W W W W W D D D D C C C C R R R R W W W W W W W W
CpG site
Increases with age
Decreases with age
4:120011960
9:121707995
17:46522122
12:108353155
11:61342647
2:25163672
15:101930774
2:156224117
18:37766939
9:70322357
2:166246798
9:119972340
11:101133279
7:67562572
8:3600628
2:169700646
10:80971848
10:80971894
4:117342269
17:56154083
2 months
22 months

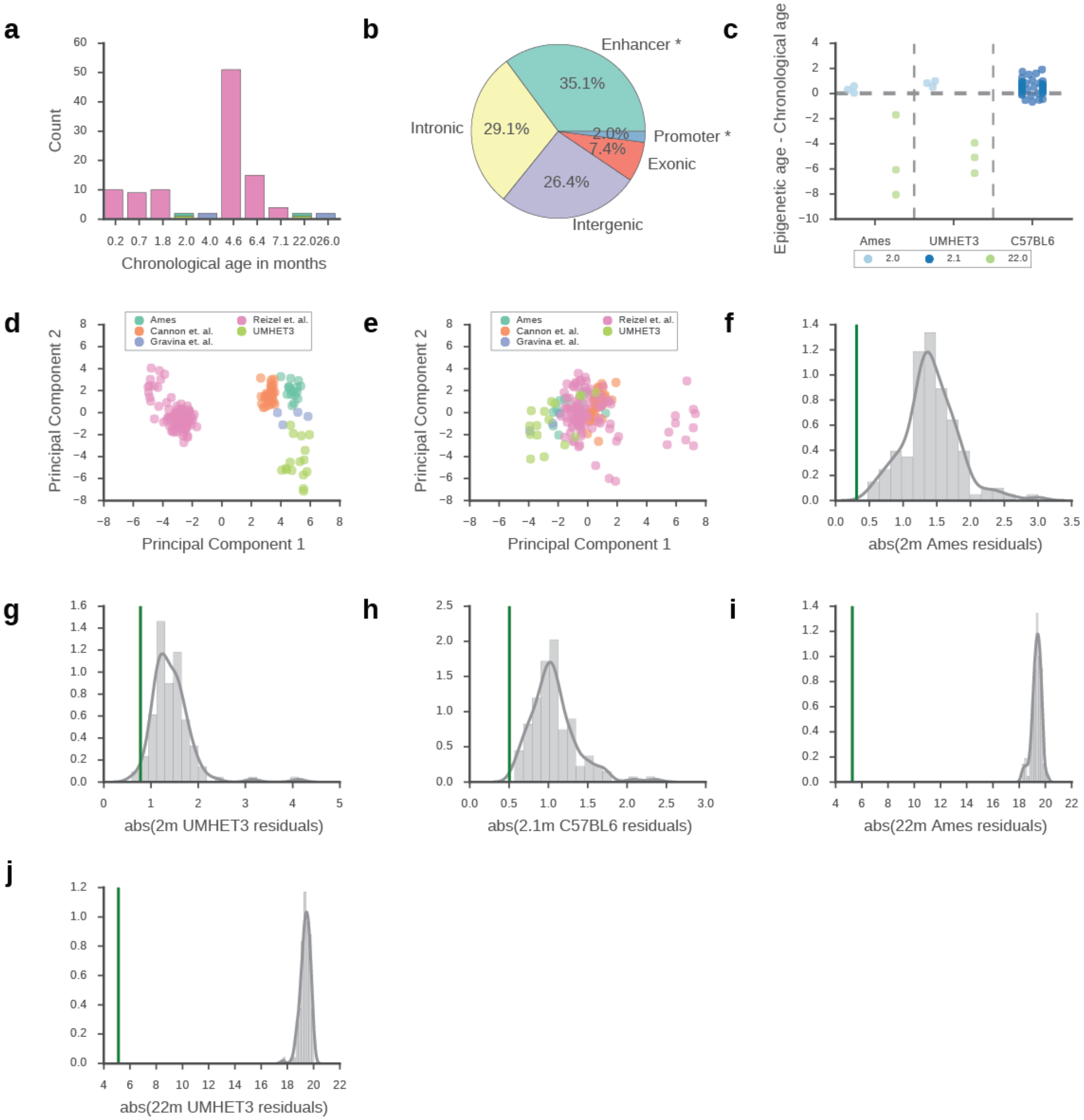
2.8 Supplementary figures

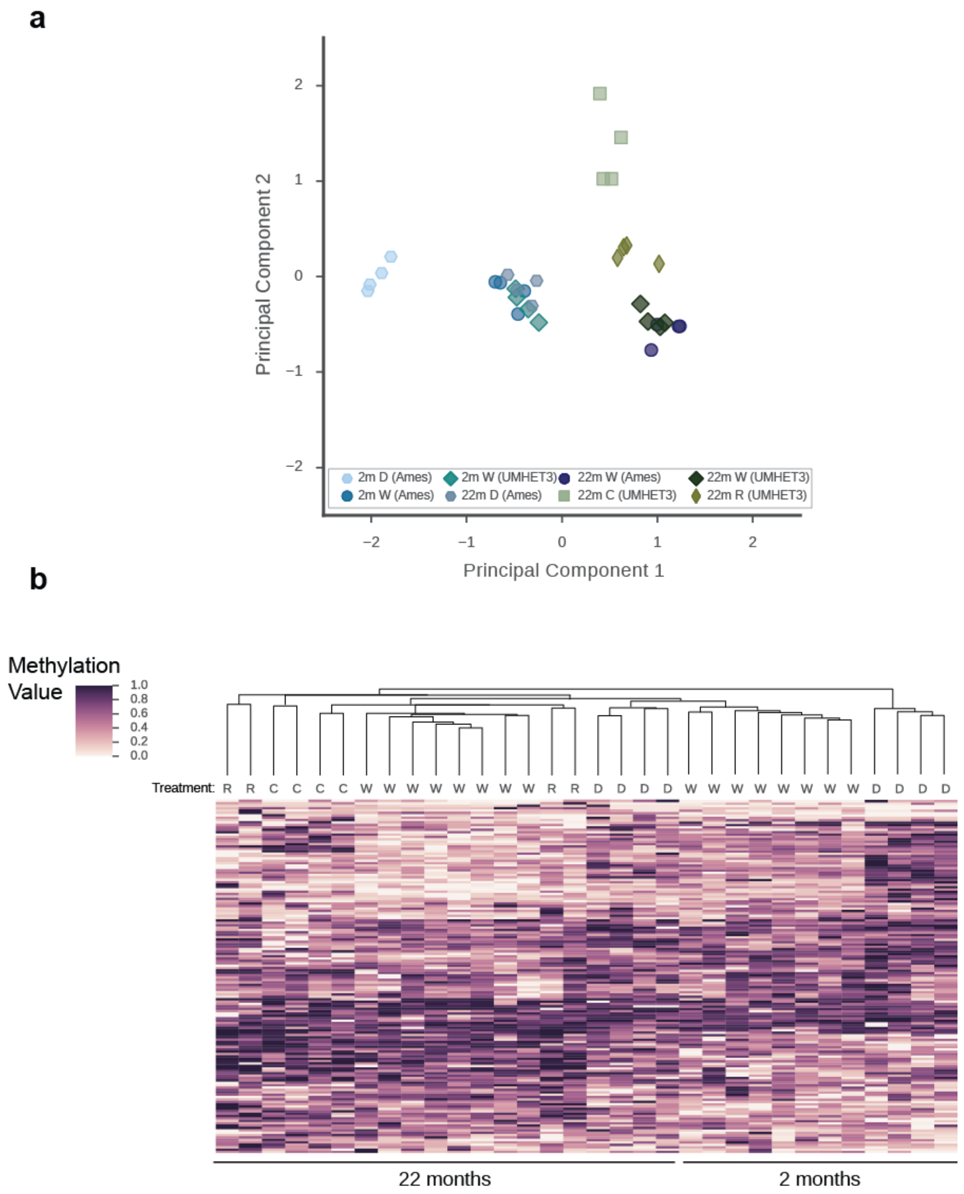


Supplmental Figure S2.1: Patterns of genomic regions affected by age-associated CpG sites Odds ratios showing enrichment (OR > 1) or depletion (OR < 1) of age-associated CpG sites in the orthologous profiled space for different genomic feature annotations (regions) in mice (left) or humans (right). * indicates p < 0.005.

Supplementary Figure S2.2: Details of data processing, model sites and model quality control.

a) Total counts of the number of mice according to their chronological ages in months used to train an ElasticNet regression (107 total). Colors correspond to sequencing studies, as shown in (**d, e**). **b)** The genomic features associated with the 148 CpG sites used in the model. There was an under-representation of these CpG sites in promoters and over-representation in enhancer regions. * indicates p < 0.01 by Fisher's exact test. **c)** Residual errors (epigenetic age - chronological age) of wild type, untreated mice of our own study and mice of Cannon *et al.* There was no detected difference in epigenetic ages across various mouse strains. Colors correspond to chronological age in months. **d)** Principal component analysis (PCA) of the 7,628 CpG sites in 173 mouse samples considered across all studies before ComBat normalization. Colors indicate the specific sequencing studies. **e)** PCA of the 7,628 CpG sites in 173 mouse samples considered across all studies after ComBat normalization, colors correspond to sequencing studies and are identical to those in (**a, d**). **f-j)** Show the results of randomizing the assignment of covariates within each study before normalizing with ComBat. After permutation and normalization, models are trained to predict age. For each permutation, models learned are then tested on the untreated, wild type mice from our study and the mice of Cannon *et al*. Residual errors are calculated for each permutation and averaged according to age and mouse genetic background. The gray bars show the distribution of average residual errors for the randomizations, where the green line indicates the residual error from the model learned on actual data. **f)** The absolute average residual error of 2-month Ames wild type mice **g)** of 2-month untreated, wild type UM-HET3 mice, **h)** of 2.1-month C57BL/6 mice, **i)** of 22-month Ames wild type mice, **j)** of 22-month untreated, wild type UM-HET3 mice. m: months


a
Count
0
10
20
30
40
50
60
0.2
0.7
1.8
2.0
4.0
4.6
6.4
7.1
22.0
26.0
Chronological age in months
b
Enhancer *
35.1%
Intronic
29.1%
2.0%
Promoter *
7.4%
Exonic
26.4%
Intergenic
c
Epigenetic age - Chronological age
4
2
0
−2
−4
−6
−8
−10
Ames
UMHET3
C57BL6
2.0
2.1
22.0
d
Ames
Cannon et. al.
Gravina et. al.
Reizel et. al.
UMHET3
Principal Component 2
Principal Component 1
−8
−6
−4
−2
0
2
4
6
8
e
Ames
Cannon et. al.
Gravina et. al.
Reizel et. al.
UMHET3
Principal Component 2
Principal Component 1
f
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.5
1.5
2.0
2.5
3.0
3.5
abs(2m Ames residuals)
g
1.6
0
1
2
3
4
5
abs(2m UMHET3 residuals)
h
0.0
0.5
1.0
1.5
2.0
2.5
3.0
abs(2.1m C57BL6 residuals)
i
4
6
8
10
12
14
16
18
20
22
abs(22m Ames residuals)
j
abs(22m UMHET3 residuals)



Supplementary Figure S2.3: Unsupervised analysis of CpG sites used in epigenetic age predictor

a) PCA of all 148 CpG sites used in ElasticNet markers is shown for the first two principal components. The colors and shapes refer to the age, treatment and mouse strain indicated in the legend. **b)** Hierarchical clustering is performed identically as performed in **Figure 2.3E**, except for all 148 CpG markers. Treatments are labeled beneath the dendogram and the ages are indicated at the bottom of the heatmap. m: months, R: rapamycin treated, C: calorie restricted, W: untreated, wild type Ames or UM-HET3

2.9 Supplementary tables and files

Supplementary File S2.1: Description of datasets used.
This file shows the table of all datasets considered for this study and various sequencing stats. It is organized into three tabs and called out accordingly in the chapter.

Supplementary File S2.2: Sites used for mouse age model.
This file shows the table of the CpGs and their weights used in the model to measure age in mice.

Supplementary File S2.3: Treatment vs wild type stats.
This file shows a series of tables of the age predictions obtained for mice treated with longevity-promoting interventions and their control/wild-type counter parts.

Supplementary Table S2.1: The effects of age and treatments on the variance of PC1

UM-HET3 Mice	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.524	0.075	-6.952	1.53E-05	-0.688 -0.360
Treatment (Calorie restriction)	-0.4656	0.097	-4.825	4.16E-04	-0.676 -0.255
Treatment (Rapamycin)	-0.2269	0.097	-2.351	3.66E-02	-0.437 -0.017
Age (months)	0.0674	0.005	13.959	8.81875E-09	0.057 0.078

Adjusted R2: 0.94

Pvalue (F statistic): 5E-08

Ames Mice	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-1.0432	0.035	-29.726	9.61E-08	-1.129 -0.957
Treatment (Dwarfism)	-1.0432	0.035	-29.726	9.61E-08	-1.129 -0.957
Age (months)	0.0764	0.004	17.004	2.65E-06	0.065 0.087

Adjusted R2: 0.98

Pvalue (F statistic): 3 E-06

This table shows the results of the effect of longevity treatment on principal component 1 assessed using regression models.

Supplementary Table S2.2: Weights of long-lived and wild type control mice used in this study

Ames mice

Genotype	Age	Min - max weight (grams)	Average
dwarf	2	6.1-7.9	7.2
wildtype	2	20.7-25.8	23
dwarf	22	15.3-28.2	23.3
wildtype	22	22.1-39.3	33.6

UM-HET3 mice

Treatment	Age	Min - max weight (grams)	Average
untreated	2	21.9-24.1	23.2
untreated	22	29.8-47.1	35.6
calorie restricted	22	19.7-23	21.2
rapamycin treated	22	29.8-47.1	35.6

2.10 Author contributions

TW and BT gathered publicly available data, performed bisulfite sequencing alignment and performed statistical analysis. NAR provided annotation files and performed bisulfite-sequencing alignment, AMG provided analysis framework for human methylome data. TW, TI, JFK, HC, PDA and HMBB designed the experiment and wrote the manuscript. All authors read and approved the final manuscript.

2.11 Acknowledgements

This work is supported by the National Institutes of health: P50 GM085764 and GCIR-06673-B awarded to TI, AG034206 and AG038509 awarded to HMBB, P01 AG031862 awarded to PDA, and T32 GM008666 awarded to TW.

We thank Richard A. Miller for providing tissue for the UM-HET3 mice and providing critical reading in the development of this manuscript.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Evidence for a common evolutionary rate in metazoan transcriptional networks" in *Genome Biology*, 2017 by Tina Wang, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

2.12 References

Ahrens, Markus, Ole Ammerpohl, Witigo von Schönfels, Julia Kolarova, Susanne Bens, Timo Itzel, Andreas Teufel, Alexander Herrmann, Mario Brosch, Holger Hinrichsen, Wiebke Erhart, Jan Egberts, Bence Sipos, Stefan Schreiber, Robert Häsler, Felix Stickel, Thomas Becker, Michael Krawczak, Christoph Röcken, Reiner Siebert, Clemens Schafmayer, and Jochen Hampe. 2013. "DNA Methylation Analysis in Nonalcoholic Fatty Liver Disease Suggests Distinct Disease-Specific and Remodeling Signatures after Bariatric Surgery." *Cell*

Metabolism 18 (2): 296–302.

Arum, Oge, Zachary Andrew Rasche, Dustin John Rickman, and Andrzej Bartke. 2013. “Prevention of Neuromusculoskeletal Frailty in Slow-Aging Ames Dwarf Mice: Longitudinal Investigation of Interaction of Longevity Genes and Caloric Restriction.” *PloS One* 8 (10): e72255.

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. “Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays.” *Bioinformatics* 30 (10): 1363–69.

Avrahami, Dana, Changhong Li, Jia Zhang, Jonathan Schug, Ran Avrahami, Shilpa Rao, Michael B. Stadler, Lukas Burger, Dirk Schübeler, Benjamin Glaser, and Klaus H. Kaestner. 2015. “Aging-Dependent Demethylation of Regulatory Elements Correlates with Chromatin State and Improved β Cell Function.” *Cell Metabolism* 22 (4): 619–32.

“Babraham Bioinformatics - Trim Galore!” 2016. Accessed December 17. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Barrett, T., S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets--Update.” *Nucleic Acids Research* 41 (Database issue): D991–95.

Bartke, Andrzej, and Holly Brown-Borg. 2004. “Life Extension in the Dwarf Mouse.” *Current Topics in Developmental Biology* 63: 189–225.

Bartke, Andrzej, and Reyhan Westbrook. 2012. “Metabolic Characteristics of Long-Lived Mice.” *Frontiers in Genetics* 3 (December): 288.

Beerman, Isabel, Christoph Bock, Brian S. Garrison, Zachary D. Smith, Hongcang Gu, Alexander Meissner, and Derrick J. Rossi. 2013. “Proliferation-Dependent Alterations of the DNA Methylation Landscape Underlie Hematopoietic Stem Cell Aging.” *Cell Stem Cell* 12 (4): 413–25.

Brown-Borg, Holly M., Kurt E. Borg, Charles J. Meliska, Andrzej Bartke, and Others. 1996. “Dwarf Mice and the Aging Process.” *Nature* 384 (6604). Macmillan Publishers Ltd., London, England: 33–33.

Cannon, Matthew V., David A. Buchner, James Hester, Hadley Miller, Ephraim Sehayek, Joseph H. Nadeau, and David Serre. 2014. “Maternal Nutrition Induces Pervasive Gene Expression Changes but No Detectable DNA Methylation Differences in the Liver of Adult Offspring.” *PloS One* 9 (3): e90335.

Gravina, Silvia, Xiao Dong, Bo Yu, and Jan Vijg. 2016. “Single-Cell Genome-Wide Bisulfite Sequencing Uncovers Extensive Heterogeneity in the Mouse Liver Methylome.” *Genome Biology* 17 (1): 150.

Gross, A. M., P. A. Jaeger, J. F. Kreisberg, K. Licon, K. L. Jepsen, M. Khosroheidari, B. M.

Morsey, S. Swindells, H. Shen, C. T. Ng, K. Flagg, D. Chen, K. Zhang, H. S. Fox, and T. Ideker. 2016. "Methylome-Wide Analysis of Chronic HIV Infection Reveals Five-Year Increase in Biological Age and Epigenetic Targeting of HLA." *Molecular Cell* 62 (2): 157–68.

Hannum, G., J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J. B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Molecular Cell* 49 (2): 359–67.

Harrison, David E., Randy Strong, Zelton Dave Sharp, James F. Nelson, Clinton M. Astle, Kevin Flurkey, Nancy L. Nadon, J. Erby Wilkinson, Krystyna Frenkel, Christy S. Carter, Marco Pahor, Martin A. Javors, Elizabeth Fernandez, and Richard A. Miller. 2009. "Rapamycin Fed Late in Life Extends Lifespan in Genetically Heterogeneous Mice." *Nature* 460 (7253): 392–95.

Hastie, Trevor, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. 2011. "Impute: Impute: Imputation for Microarray Data." R package version.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic Acids Research* 34 (Database issue): D590–98.

Horvath, S. 2013. "DNA Methylation Age of Human Tissues and Cell Types." *Genome Biology* 14 (10): R115.

Horvath, S., W. Erhart, M. Brosch, O. Ammerpohl, W. von Schonfels, M. Ahrens, N. Heits, J. T. Bell, P. C. Tsai, T. D. Spector, P. Deloukas, R. Siebert, B. Sipos, T. Becker, C. Rocken, C. Schafmayer, and J. Hampe. 2014. "Obesity Accelerates Epigenetic Aging of Human Liver." *Proceedings of the National Academy of Sciences of the United States of America* 111 (43): 15538–43.

Horvath, Steve, Paolo Garagnani, Maria Giulia Bacalini, Chiara Pirazzini, Stefano Salvioli, Davide Gentilini, Anna Maria Di Blasio, Cristina Giuliani, Spencer Tung, Harry V. Vinters, and Claudio Franceschi. 2015. "Accelerated Epigenetic Aging in Down Syndrome." *Aging Cell* 14 (3): 491–95.

Johansson, Asa, Stefan Enroth, and Ulf Gyllensten. 2013. "Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan." *PloS One* 8 (6): e67378.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics* 8 (1): 118–27.

Jones, Eric, Travis Oliphant, Pearu Peterson, and Others. 2015. "SciPy: Open Source Scientific Tools for Python, 2001." *URL Http://www. Scipy. Org* 73: 86.

Krueger, Felix, and Simon R. Andrews. 2011. “Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications.” *Bioinformatics* 27 (11): 1571–72.

Laird, Peter W. 2010. “Principles and Challenges of Genome-Wide DNA Methylation Analysis.” *Nature Reviews. Genetics* 11 (3). Nature Publishing Group: 191–203.

Lamming, Dudley W., Lan Ye, Pekka Katajisto, Marcus D. Goncalves, Maki Saitoh, Deanna M. Stevens, James G. Davis, Adam B. Salmon, Arlan Richardson, Rexford S. Ahima, David A. Guertin, David M. Sabatini, and Joseph A. Baur. 2012. “Rapamycin-Induced Insulin Resistance Is Mediated by mTORC2 Loss and Uncoupled from Longevity.” *Science* 335 (6076): 1638–43.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.

Leek, Jeffrey T., W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. 2012. “The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments.” *Bioinformatics* 28 (6): 882–83.

Lopez-Otin, C., M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. 2013. “The Hallmarks of Aging.” *Cell* 153 (6): 1194–1217.

Maegawa, Shinji, Sheryl M. Gough, Naoko Watanabe-Okochi, Yue Lu, Nianxiang Zhang, Ryan J. Castoro, Marcos R. H. Estecio, Jaroslav Jelinek, Shoudan Liang, Toshio Kitamura, Peter D. Aplan, and Jean-Pierre J. Issa. 2014. “Age-Related Epigenetic Drift in the Pathogenesis of MDS and AML.” *Genome Research* 24 (4): 580–91.

Marioni, R. E., S. Shah, A. F. McRae, B. H. Chen, E. Colicino, S. E. Harris, J. Gibson, A. K. Henders, P. Redmond, S. R. Cox, A. Pattie, J. Corley, L. Murphy, N. G. Martin, G. W. Montgomery, A. P. Feinberg, M. D. Fallin, M. L. Multhaup, A. E. Jaffe, R. Joehanes, J. Schwartz, A. C. Just, K. L. Lunetta, J. M. Murabito, J. M. Starr, S. Horvath, A. A. Baccarelli, D. Levy, P. M. Visscher, N. R. Wray, and I. J. Deary. 2015. “DNA Methylation Age of Blood Predicts All-Cause Mortality in Later Life.” *Genome Biology* 16: 25.

Masternak, Michal M., Jacob A. Panici, Michael S. Bonkowski, Larry F. Hughes, and Andrzej Bartke. 2009. “Insulin Sensitivity as a Key Mediator of Growth Hormone Actions on Longevity.” *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 64 (5): 516–21.

McCay, Cmw, Mary F. Crowell, L. A. Maynard, and Others. 1935. “The Effect of Retarded Growth upon the Length of Life Span and upon the Ultimate Body Size.” *The Journal of Nutrition* 10 (1): 63–79.

Means, L. W., J. L. Higgins, and T. J. Fernandez. 1993. “Mid-Life Onset of Dietary Restriction Extends Life and Prolongs Cognitive Functioning.” *Physiology & Behavior* 54 (3): 503–8.

Medvedev, Z. A. 1990. "An Attempt at a Rational Classification of Theories of Ageing." *Biological Reviews of the Cambridge Philosophical Society* 65 (3): 375–98.

Miller, Richard A., David E. Harrison, Clinton M. Astle, Elizabeth Fernandez, Kevin Flurkey, Melissa Han, Martin A. Javors, Xinna Li, Nancy L. Nadon, James F. Nelson, and Others. 2014. "Rapamycin-Mediated Lifespan Increase in Mice Is Dose and Sex Dependent and Metabolically Distinct from Dietary Restriction." *Aging Cell* 13 (3). Wiley Online Library: 468–77.

Orozco, Luz D., Marco Morselli, Liudmilla Rubbi, Weilong Guo, James Go, Huwenbo Shi, David Lopez, Nicholas A. Furlotte, Brian J. Bennett, Charles R. Farber, Anatole Ghazalpour, Michael Q. Zhang, Renata Bahous, Rima Rozen, Aldons J. Lusis, and Matteo Pellegrini. 2015. "Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice." *Cell Metabolism* 21 (6): 905–17.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (November). JMLR.org: 2825–30.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

Reizel, Yitzhak, Adam Spiro, Ofra Sabag, Yael Skversky, Merav Hecht, Ilana Keshet, Benjamin P. Berman, and Howard Cedar. 2015. "Gender-Specific Postnatal Demethylation and Establishment of Epigenetic Memory." *Genes & Development* 29 (9): 923–33.

Rosenbloom, Kate R., Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Glenn Hickey, Angie S. Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T. Lee, Chin H. Li, Karen H. Miga, Ngan Nguyen, Benedict Paten, Brian J. Raney, Arian F. A. Smit, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. 2015. "The UCSC Genome Browser Database: 2015 Update." *Nucleic Acids Research* 43 (Database issue): D670–81.

Rosenbloom, Kate R., Cricket A. Sloan, Venkat S. Malladi, Timothy R. Dreszer, Katrina Learned, Vanessa M. Kirkup, Matthew C. Wong, Morgan Maddren, Ruihua Fang, Steven G. Heitner, Brian T. Lee, Galt P. Barber, Rachel A. Harte, Mark Diekhans, Jeffrey C. Long, Steven P. Wilder, Ann S. Zweig, Donna Karolchik, Robert M. Kuhn, David Haussler, and W. James Kent. 2013. "ENCODE Data in the UCSC Genome Browser: Year 5 Update." *Nucleic Acids Research* 41 (Database issue): D56–63.

Spiers, H., E. Hannon, S. Wells, B. Williams, C. Fernandes, and J. Mill. 2016. "Age-Associated Changes in DNA Methylation across Multiple Tissues in an Inbred Mouse Model." *Mechanisms of Ageing and Development* 154: 20–23.

Sun, Deqiang, Min Luo, Mira Jeong, Benjamin Rodriguez, Zheng Xia, Rebecca Hannah, Hui Wang, Thuc Le, Kym F. Faull, Rui Chen, Hongcang Gu, Christoph Bock, Alexander

Meissner, Berthold Göttgens, Gretchen J. Darlington, Wei Li, and Margaret A. Goodell. 2014. "Epigenomic Profiling of Young and Aged HSCs Reveals Concerted Changes during Aging That Reinforce Self-Renewal." *Cell Stem Cell* 14 (5): 673–88.

Takasugi, Masaki. 2011. "Progressive Age-Dependent DNA Methylation Changes Start before Adulthood in Mouse Tissues." *Mechanisms of Ageing and Development* 132 (1-2): 65–71.

Teschendorff, Andrew E., Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. "A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 K DNA Methylation Data." *Bioinformatics* 29 (2): 189–96.

Urich, Mark A., Joseph R. Nery, Ryan Lister, Robert J. Schmitz, and Joseph R. Ecker. 2015. "MethylC-Seq Library Preparation for Base-Resolution Whole-Genome Bisulfite Sequencing." *Nature Protocols* 10 (3): 475–83.

Waskom, Michael, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Alistair Miles, Tom Augspurger, Tal Yarkoni, Tobias Megies, Luis Pedro Coelho, Daniel Wehner, cynddl, Erik Ziegler, diego, Yury V. Zaytsev, Travis Hoppe, Skipper Seabold, Phillip Cloud, Miikka Koskinen, Kyle Meyer, Adel Qalieh, and Dan Allan. 2014. *Seaborn: v0.5.0 (November 2014)*. ZENODO. doi:10.5281/zenodo.12710.

Weidner, C. I., Q. Lin, C. M. Koch, L. Eisele, F. Beier, P. Ziegler, D. O. Bauerschlag, K. H. Jockel, R. Erbel, T. W. Muhleisen, M. Zenke, T. H. Brummendorf, and W. Wagner. 2014. "Aging of Blood Can Be Tracked by DNA Methylation Changes at Just Three CpG Sites." *Genome Biology* 15 (2): R24.

Wilkinson, John E., Lisa Burmeister, Susan V. Brooks, Chi-Chao Chan, Sabrina Friedline, David E. Harrison, James F. Hejtmancik, Nancy Nadon, Randy Strong, Lauren K. Wood, Maria A. Woodward, and Richard A. Miller. 2012. "Rapamycin Slows Aging in Mice." *Aging Cell* 11 (4): 675–82.

Yates, Andrew, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. 2016. "Ensembl 2016." *Nucleic Acids Research* 44 (D1): D710–16.

Zampieri, Michele, Fabio Ciccarone, Roberta Calabrese, Claudio Franceschi, Alexander Bürkle, and Paola Caiafa. 2015. "Reconfiguration of DNA Methylation in Aging." *Mechanisms of Ageing and Development* 151 (November): 60–70.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2). Blackwell

Publishing Ltd: 301–20.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2): 301–20.

CHAPTER 3: A conserved epigenetic progression aligns human and dog age

3.1 Abstract

Mammals progress through common physiological stages as they age, from early development to puberty, senescence, and death (Kirkwood 2005). Yet the degree to which these physiological changes reflect a common molecular program is unclear (Cohen 2018), making it difficult to study or potentially modify the mechanisms of aging (Kirkwood 2005; Cohen 2018). Here we chart the conserved impacts of aging on the mammalian genome, focusing on evolutionary comparisons of humans with dogs, a compelling emerging model for aging (Gilmore and Greer 2015; Kaeberlein, Creevy, and Promislow 2016). Using a strategy we call syntenic bisulfite sequencing, we profile the methylomes of 104 Labrador retrievers of diverse ages, achieving >150X coverage within mammalian syntenic blocks. Correlation with human methylomes reveals that the two species experience a common progression of epigenetic changes during the course of life. The progression appears highly non-linear, with a very rapid transformation in puppies relative to children which slows markedly in adulthood. Based on these results, we estimate a logarithmic function for epigenetic translation of dog to human years. This function correctly translates the times at which major physiological milestones of aging are observed in the two species, and it extends to methylation patterns in mice, suggesting that conserved epigenetic signals may underlie common aging physiology in many mammals. We find these results are driven by CpG methylation marks near genes functioning during development, and that these sites are sufficient to translate biological age across species. Our results identify a universal epigenetic clock that underlies mammalian aging, opening the door to comparative studies of aging and aging interventions in diverse organisms.

3.2 Introduction

Despite large differences in lifespan, mammals nonetheless undergo a similar progression of changes in physiology as they age, from sexual maturation until death (Kirkwood 2005). Such conserved aging physiology is perplexing, since age-related traits acquired after sexual maturity are not necessarily constrained by evolution (Kirkwood 2005). As a consequence, aging mechanisms in short-lived mammals might differ from those in long-lived mammals (Cohen 2018) which, if generally true, would substantially complicate efforts to understand the molecular basis of aging.

DNA methylation, an epigenetic modification frequently found at cytosine-guanine dinucleotides (CpGs), may provide insight into how different species age (Field et al. 2018). Changes in DNA methylation have been strongly associated with the process of human aging(Field et al. 2018; Lopez-Otin et al. 2013), such that the methylation states of a few well-chosen CpGs can be used to recover an individual's chronological age if unknown (Vidaki et al. 2017), or to estimate a relative rate of aging (Gross et al. 2016; Lu et al. 2019). Such "epigenetic clocks" (Hannum et al. 2013; Horvath 2013; T. Wang et al. 2017; Petkovich et al. 2017; Thompson et al. 2017; Stubbs et al. 2017) have also been demonstrated in various other mammals using CpG markers selected independently for each species in separate studies. Thus it is clear that aging remodels the methylome, and that it does so in multiple species. However, to what extent this epigenetic remodeling is conserved, or largely species specific, remains an open question (Stubbs et al. 2017; T. Wang et al. 2017).

Domestic dogs provide a unique opportunity to investigate this question (Kaeberlein, Creevy, and Promislow 2016; Gilmore and Greer 2015). Dogs have been selectively bred by humans for occupation and aesthetics (Ostrander et al. 2017), resulting in breeds within which members share characteristic traits. Each breed was carefully established using small numbers of founders, thus intrabreed genetic variation is lower than that observed between even closely related breeds (Dreger et al. 2016). For these reasons, small numbers of dogs from the same or related breeds can be used to decipher the genetic basis of complex traits (Davis and Ostrander 2014) such as aging. Although humans and dogs diverged early during mammalian evolution (Vonholdt et al. 2010), dogs exhibit similar age-related pathologies as humans in a much shorter lifespan(Kaeberlein, Creevy, and Promislow 2016; Gilmore and Greer 2015). Moreover, dogs share nearly all aspects of their environment with humans, including similar levels of health care (Kaeberlein, Creevy, and Promislow 2016; Gilmore and Greer 2015). Together, these facts have established dogs as an important system for comparative studies of aging and exploration of lifespan-enhancing interventions, leading to prominent consortia such as the Dog Aging Project (Kaeberlein, Creevy, and Promislow 2016).

3.3 Results

To enable comparisons of epigenetic aging in dogs and other mammals, we developed a strategy to measure DNA methylation within regions of synteny across mammalian genomes (**Figure 3.1a**). By this strategy, henceforth called *Synteny Bisulfite Sequencing* (SyBS), oligonucleotide probes were designed to capture approximately one million CpGs in mammalian syntenic regions (**Methods**). We then applied SyBS to characterize the methylomes of 104 dogs, primarily consisting of Labrador retrievers representing the entire lifespan, from 0.1 to 16 years at the time of blood draw (**Supplementary Figure S3.1a**, **Supplementary Table 3.1**). Libraries were sequenced to an average depth of 163X, with nine dogs removed due to lack of coverage. The methylation values of captured CpGs were similar to those obtained using whole-genome bisulfite sequencing (Pearson's correlation $r > 0.85$, **Supplementary Figure S3.1b**) and across independent captures from the same DNA samples ($r > 0.95$, **Supplementary Figure S3.1c-h**). As expected by design, SyBS achieved approximately 13-times higher coverage of syntenic regions compared to non-targeted bisulfite methods (**Figure 3.1b**). Captured CpG sites were located throughout the genome and were enriched for exons and CpG islands (**Figure 3.1c**). Such regions were also among those targeted by the Illumina methylome arrays frequently used in human methylation studies (Dedeurwaerder et al. 2011). Accordingly, we obtained previously published methylation profiles from the blood of 320 human individuals aged 1 to 103 years at the time of sample isolation (Alisch et al. 2012; Hannum et al. 2013). Based on these data, we identified 54,469 well-profiled CpGs that were orthologous in dogs and humans, thereby enabling systematic evolutionary studies of epigenetic aging (**Methods**).

Comparing these methylome-wide profiles across species (Pearson's correlation, **Methods**), we observed the highest similarities when pairing dogs and humans of the same relative ages, such as young *versus* young or old *versus* old individuals, and the lowest similarities when pairing young dogs with old humans and *vice versa* (**Figure 3.2a**). This relationship between methylome similarity and age was abolished upon permutation (FDR < 0.01; **Supplementary Figure S3.2a**), suggesting the presence of a conserved progression of epigenetic changes during dog and human aging. Notably, this signal was sufficiently strong to arise in a genome-wide unsupervised analysis without any specific subselection of markers.

An immediate question was whether the epigenetic progression was merely binary, representing a young and an old biological state, or could be seen to translate age across species with greater time resolution. To address this question, we assigned the age of each dog to the average age of its nearest *k* humans by methylome-wide similarity (**Methods**). This analysis revealed a monotonic time-resolved relationship between dog and human age which was however highly nonlinear (**Figure 3.2b** for k = 5, **Supplementary Figure S3.2f-g** for other k). Similar results were obtained in a reciprocal analysis assigning each human to its nearest dogs (**Figure 3.2c**), prompting us to combine the two reciprocal analyses to fit the single function $human\_age = 16\ ln(dog\_age) + 31$ (**Figure 3.2d**).

Although this logarithmic function was a significant departure from the conventional wisdom that one year of a dog's life equates to seven human years, we noted better agreement with the approximate times at which dogs and humans experience common aging physiology (infant, juvenile, adolescent, mature, senior) (Lebeau 1953; Bogin and Smith 1996; Bartges et al. 2012) (**Figure 3.2d**). The agreement between epigenetics and aging physiology was particularly close for infant, juvenile and senior stages: For instance, the epigenome translated 7 weeks (0.15 years) in dogs to 9 months (0.78 years) in humans, corresponding to the infant stage when deciduous teeth erupt in both puppies and babies(Bogin and Smith 1996; Bartges et al. 2012). In seniors, the current expected lifespan of Labrador retrievers, 12 years, correctly translated to the current worldwide lifetime expectancy of humans, 70 years (Cia 2013; Fleming, Creevy, and Promislow 2011). For adolescent and mature stages, the correspondence was more approximate, with the epigenome showing faster aging for dogs, relative to humans, than expected by physiological tables (Inoue et al. 2015; Arias, Heron, and Xu 2017) (**Figure 3.2d**). Thus, the epigenome progresses through a series of conserved biological states that inform the major physiological changes of aging, which for dogs and humans occur in the same sequence but at markedly different times relative to total lifespan.

Further support for a conserved epigenetic program was obtained through comparisons with the methylomes of 133 mice, aged 3 months to 3 years (Petkovich et al. 2017). As with the dog-human comparisons, we found that dog and mouse methylomes were most similar for individuals in similar age quantiles (**Supplementary Figure S3.3, Methods**). The ability to detect a conserved epigenetic progression in a third mammal suggested this progression may be fundamental to aging in many mammalian species.

To determine whether the conserved changes were concentrated within particular genes or gene functions, we examined CpG methylation states near 7,942 genes for which orthologs were clearly present in all three species: dogs, humans, and mice (**Methods**). This analysis identified 394 genes for which methylation values showed conserved aging behavior, i.e. increases or decreases with age that were coherent across species (empirical $p < 0.05$, **Supplementary Figure S3.4**). To understand these gene functions we mapped them onto the Parsimonious Composite Network (PCNet), a large repository of approximately 2×10^6 molecular interactions representing physical and functional relationships among genes and gene products in which each interaction has support from multiple sources (Huang et al. 2018). Genes clustered into five highly interconnected network modules (**Figure 3.3**). Four were enriched for distinct developmental functions in anatomical development (2 modules, 117 and 69 genes), synapse assembly (18) and neuroepithelial cell differentiation (5), with the majority of the genes in these developmental pathways increasing in methylation with age ($FDR < 0.05$). In contrast, the remaining pathway module consisted of 144 genes that were enriched in leukocyte differentiation and broad metabolic functions (RNA metabolism, response to organic substrates) and had primarily decreasing methylation values with age. Notably, genes that increase methylation with age were also among those that exhibited conserved age-related expression changes in humans and mice (Skene, Roy, and Grant 2017) (Observed/expected = 2.5, $p < 0.01$ by hypergeometric test). Strikingly, the developmental genes in these five modules were among the most highly conserved in DNA sequence in the mammalian genome, even accounting for the already high sequence conservation of developmental genes in general (**Figure S3.5**).

As further assessment of the importance of developmental pathways in the conserved epigenetic progression, we recalculated methylome similarity across species using only methylation changes at developmental genes. The revised analysis yielded cross-species translations of age that were very similar to those observed earlier when using all CpGs (**Supplementary Figure S3.6a**). In contrast, such translations were substantially degraded when removing CpGs near developmental genes from the analysis (**Supplementary Figure S3.6b-f**). Thus, methylation changes near developmental genes appeared to be both necessary and sufficient for the conserved epigenetic progression observed during mammalian life.

If the methylation status of developmental modules indeed tracks the physiological progression of the organism and not just chronological time, an important prediction is that it will respond to interventions that slow or delay this physiology, such as anti-aging treatments. In mice, calorie restriction and dwarfism have been associated with increased lifespan relative to control animals (Petkovich et al. 2017). We therefore examined whether such effects could be seen within the conserved development gene modules (as identified in **Figure 3.5**). For this purpose we constructed an epigenetic clock, a regularized linear regression model that measures age from CpG methylation values (Hannum et al. 2013; Horvath 2013; T. Wang et al. 2017; Petkovich et al. 2017; Thompson et al. 2017; Stubbs et al. 2017), focusing exclusively on 439 developmental module CpGs (**Figure 3.4a**, **Methods**). By training in dogs or alternatively in mice, this "conserved development" clock could be used to translate a methylation profile either to dog or mouse years (**Figures 3.4b-c**), including cross-translation of a dog methylome to its equivalent mouse age or *vice versa* (**Figure 3.4d**). In fact, epigenetic ages measured by this clock were more consistent with actual ages of those animals than clock measurements formulated from the rest of the methylome (**Figure 3.4d**). When applying the conserved development clock to mice treated with lifespan-extending interventions, the measured epigenetic ages were significantly less than those of control mice, by 30% on average ($p < 10^{-6}$, **Figure 3.4e**). These results were also clearly observed when using the conserved development clock trained in dogs to predict mouse age (**Figure 3.4f**). Together, these results demonstrate that the methylation states of developmental gene modules track the physiological effects of aging and aging interventions in multiple mammalian species.

3.4 Discussion

In summary, we have found that methylation patterns follow a conserved epigenetic progression during mammalian aging. Using this progression, we have identified a conserved clock that translates both chronological age and biological age from one species to another, despite large differences in lifespan. These results demonstrate that the physiological signs of aging are embedded into the mammalian epigenome. Notably, the conserved epigenetic progression predominantly affected highly sequence-constrained developmental genes, similar to a previous study using highly-constrained ribosomal DNA (M. Wang and Lemos 2019). These findings suggest that this progression may be determined, in part, by the sequence of these genes, or perhaps other conserved mechanisms affecting highly constrained genes. Identifying factors that slow this progression in tractable, short-lived model species provides an exciting route to understand the mechanisms of human aging and their potential interventions.

3.5 Methods

Annotations. Reference genomes were downloaded from Ensembl for dog (CanFam3.1), mouse (mm10) and human (hg19). Ensembl Biomart version 91 was used for gene, 3'UTR and

5’UTR annotations (Yates et al. 2016). CpG islands, repeat annotations, and chain files were downloaded from the UCSC Genome Browser (Rosenbloom et al. 2015). CpG shores were designated as regions 2 kilobases (kb) outside each CpG island, and CpG shelves were designated as regions 2kb outside of CpG shores. Promoters were designated as regions 2kb upstream and 100 basepairs (bp) downstream of the transcription start sites (TSS) based on gene annotations from Ensembl (Yates et al. 2016). Whole genes were divided into exonic and intronic sequences. Intergenic regions were then defined as the remaining regions of the genome after subtracting all other annotated regions. Definitions of one-to-one orthologs were downloaded from Ensembl Compara (Vilella et al. 2009) for dogs, humans and mice.

Public Datasets. The following datasets were obtained from Gene Expression Omnibus (GEO) or Sequence Read Archives (SRA) (*source articles in parenthesis*) [number of individuals included in study in brackets]:

- GSE80672 (Petkovich et al. 2017): Methylomes from postnatal mice. Blood, Reduced Representation Bisulfite Sequencing (RRBS) method. [133]
- GSE36054 (Alisch et al. 2012): Methylomes from human children. Blood, Infinium 450K array. [35]
- GSE40279 (Hannum et al. 2013): Methylomes from human adults. Blood, Infinium 450K array. [285]
- SRP065319 (Thompson et al. 2017): Methylomes from dogs and wolves. Blood, RRBS method. [92]

Canine samples. Information on each dog sample used, including age, breed, and source, is given in **Supplementary Table 3.1**, with the age distribution also provided in **Supplementary Figure S3.1a**. For samples sourced from NHGRI, domestic dogs were collected with owners’ signed consents in accordance with standard protocols approved by the NHGRI IACUC committee. Samples were collected at canine centric events such as dog shows, obedience training sessions and competitive games. Alternatively, owners were supplied with a mail-in kit which included instructions, tubes for blood draws and a general information sheet with the AKC number (when available), pedigree and date of birth. Blood draws were performed by licensed veterinarians or veterinary technicians.

For samples sourced from UC Davis, blood samples were collected from privately owned dogs through the William R. Pritchard Veterinary Medical Teaching Hospital. Owners specified the breed of each dog. Standard collection protocols were reviewed and approved by the UC Davis IACUC.

For samples from NHGRI, DNA was extracted using a well-established cell lysis protocol described by (Bell, Karam, and Rutter 1981), followed by phenol/chloroform extraction with phase separation performed in 15 mL phase-lock tubes (5-Prime, Inc. Gaithersburg, MD, USA). For other samples, DNA was extracted using the Puregene kit (Qiagen).

SyBS Target selection. Our strategy for syntenic bisulfite sequencing was to identify highly syntenic regions among mammalian genomes and then specifically select those covered by the Illumina Human 450K methylome array, thereby enabling informative comparisons to pre-existing human methylome data generated using the Illumina system. First, highly syntenic regions were determined using PhyloP scores(Siepel, Pollard, and Haussler 2006) among placental mammals (46-vertebrate alignment obtained from UCSC genome browser)(Rosenbloom et al. 2015). We excluded regions that aligned to sex chromosomes in dogs. Hybridization probes were generated to target these regions using the Roche SeqCap-Epi platform. This process resulted in an 18.8mb (megabase) sequencing library in dogs, containing approximately 90,000 CpGs that were also profiled by the Illumina 450K array in humans.

SyBS Library preparation and sequencing. We followed the protocol specified by the Roche SeqCap-Epi platform. Briefly, approximately 500ng of lambda phage DNA (bisulfite-conversion control) was added to 1ug of dog DNA, then sheared to an average of 175bp (Covaris). Sheared DNA was end-repaired, A-tailed and ligated to barcoded adapters. Adapter-ligated libraries were subjected to bisulfite treatment (Zymo EZ DNA methylation lightning kit) following manufacturer instructions. Bisulfite-treated libraries were cleaned and amplified using 25 cycles of PCR with a uracil-tolerant enzyme (Kapa). Libraries were pooled equimolarly into 4-plex or 6-plex hybridization capture reactions to a total of 1ug per reaction. Captured product was PCR amplified (10 cycles). Hybridizations were pooled before sequencing and split among 10 lanes on an Illumina HiSeq 4000 in 2x150bp cycles.

SyBS data analysis. Reads obtained from sequencing were demultiplexed and their quality was verified using FastQC (Andrews and Others 2010). Reads were trimmed using TrimGalore ("Babraham Bioinformatics - Trim Galore!" 2016) (4bp) then aligned to a bisulfite-converted dog genome (CanFam3.1) using Bismark (v0.14.3) (Krueger and Andrews 2011), which produces alignments with Bowtie2 (v2.1.0) (Langmead et al. 2009), with parameters "-score_min L,0,-0.2". Methylation values for CpG sites were determined using MethylDackel (v0.2.1)("MethylDackel," n.d.). Custom Python scripts using BEDtools (v2.25.0) (Quinlan and Hall 2010) were used to determine on-target reads. Optical PCR duplicates were determined using Picard tools (v1.141) ("Picard Tools - By Broad Institute" 2017) and removed using Samtools (v0.1.18) (Li et al. 2009). Coverage of syntenic regions was determined using the number of unique on-target reads that were orthologous to humans, divided by the expected sequencing space. Only CpG sites that were on-target, covered by at least five reads and present across 90% of samples were selected for further analysis. Samples missing more than 30% of CpGs were removed from further analyses resulting in the removal of nine dogs. Missing data for selected CpGs were imputed by performing k-nearest neighbors ($k = 10$) using fancyimpute (“Fancyimpute,” n.d.) in Python.

To assess the concordance of methylation values obtained using SyBS with conventional approaches, we also sequenced 10 dogs using whole-genome bisulfite sequencing (libraries prior to enrichment with SyBS probes). Reads were processed and aligned with the canine genome as described above. We saw an average Pearson’s correlation of $r = 0.85$ among these 10 samples (range 0.75 - 0.97) (**Supplementary Figure S3.1b**). We also performed independent replicate hybridizations for 6 samples. We saw an average $r = 0.97$ (range: 0.96 - 0.98) for these technical replicates (**Supplementary Figure S3.1c-h**). We also verified that lambda phage DNA exhibited complete conversion (>99.5%).

Public RRBS data processing. For previous data generated using Reduced Representation Bisulfite Sequencing (RRBS), methods for alignment and CpG selection were identical to those described above. Since RRBS fragments are generated using restriction enzymes with specific recognition sites, optical PCR duplicates could not be removed and on-target CpGs were not determined. For evolutionary comparative analysis, we included 133 control mice aged between 3 months to 2.5 years (Petkovich et al. 2017). To compare the coverage of syntenic regions between SyBS and non-targeted bisulfite technology, we used a RRBS study in dogs and wolves (Thompson et al. 2017) (**Figure 3.1b**).

Human methylation array data processing. Illumina Infinium 450K methylome array data were quantile normalized using Minfi (Aryee et al. 2014) and missing values were imputed using the Impute package in R (Hastie et al. 2011). These values were adjusted for cell counts as described by (Gross et al. 2016). To enable comparisons across different methylation array studies, we implemented beta-mixture quantile dilation (BMIQ) (Teschendorff et al. 2013; Horvath 2013), and used the median of the Hannum et al. dataset as the gold standard (Hannum et al. 2013). To mitigate residual batch effects, we selected human samples that clustered closely in the first two principal components using scikit-learn v0.19.2 (Pedregosa et al. 2011) and verified that such filtering had little effect on the distribution of ages. We also removed samples for which more than 10% of probes were not adequately detected. This procedure resulted in methylome profiles for 320 humans that could be compared to the SyBS generated dog methylomes.

<u>Determining orthologous CpGs.</u> Human Illumina 450K methylation array CpGs were extended by 50bp with respect to the strand using BEDtools and mapped to the target genome (mouse or dog) using liftOver with "-minMatch=0.5". We verified that the coordinate alignment obtained using 50bp was identical to that obtained using the exact coordinate (1bp) at "-minMatch=0.95". This procedure allowed us to determine an exact orthologous region for each human CpG and each dog CpG. When multiple dog CpGs were assigned to one human CpG probe region, we took the average methylation value of the aligned CpGs in dogs. This procedure resulted in 54,469 dog-human orthologous CpGs for further analysis. To mitigate batch effects from sequencing versus array platforms, we normalized the sequencing methylation values using BMIQ and performed quantile normalization using the preprocessCore package in R (normalize.quantiles.use.target function) (Bolstad 2013).

For dog-to-mouse comparisons, CpGs that were separated by 1bp were merged into one region using BEDtools. Each region was then extended by 50bp. The resulting region files were aligned to the target genome using liftOver "-minMatch=0.5". Only regions that were concordant between the two alignments (*i.e.*, dog to mouse or mouse to dog) were selected for further analysis.

CpGs that were assigned to the same aligned regions were averaged to generate 9,404 bins (consisting of 87,915 CpGs from dogs).

Computing dog-human pairwise methylome similarity. Methylation values of orthologous CpGs were normalized by subtracting the mean and dividing by the standard deviation over individuals (i.e. z-transformed, separately for each species). The resulting z-values represent the tendency to decrease or increase relative to the mean of each CpG within a species. Using these values we calculated the pairwise Pearson's correlation between the methylomes for each dog-human pair. Correlation was computed across all orthologous CpG values using the SciPy Python package(Jones et al. 2015), forming a 95 x 320 (dog x human) methylome similarity matrix (*MS*). We also created a coarsened version of this matrix, in which the pairwise similarities were averaged over two-year age windows in both species, forming an 8 x 51 (dog x human) methylome similarity matrix (*MSA*, **Supplementary Figure S3.2a**).

Given this matrix, we evaluated the significance of association between age and methylome similarity using permutations. Specifically, we generated the following two-by-two contingency table:

	Ages more different $AD(i,j) > \underline{AD}$	Ages more similar $AD(i,j) \leq \underline{AD}$
Methylomes more different $MSA(i,j) \leq \underline{MSA}$	$Count_1$	$Count_3$
Methylomes more similar $MSA(i,j) > \underline{MSA}$	$Count_2$	$Count_4$

where MSA is the methylome similarity matrix, $AD(i,j)$ is $\left|Age\ bin_{dog} - Age\ bin_{human}\right|andCount$is the number of occurrences (cells within the *MSA* similarity matrix) for which the table row and column conditions are met. Using these counts, we calculated the p-value using the one tailed Fisher's exact test and compared this p-value to that obtained when permuting the membership of dogs and humans in two-year age bins across 1000 permutations (**Supplementary Figure S3.2a**).

<u>k-nearest neighbors analysis.</u> To achieve a robust assignment of reciprocal nearest neighbors, we used a strategy inspired by Context Likelihood of Relatedness (Madar et al. 2010). Specifically, we z-normalized the *MS* methylome similarity matrix to form MSZ, as follows:

$$MSZ_{row}(i,j) = max(\,0, \frac{MS_{i,j} - \underline{MS_{i*}}}{\sigma_{i*}})$$

$$MSZ_{column}(i,j) = max(\,0, \frac{MS_{i,j} - \underline{MS_{*j}}}{\sigma_{*j}})$$

$$MSZ(i,j) = mean[MSZ_{row}(i,j)\,, MSZ_{column}(i,j)]$$

k-nearest neighbors were assigned to each dog or to each human with respect to MSZ values. This process was implemented in Python using scikit-learn**.** We evaluated the significance by permuting either dog age or human age (1000 shuffles each, for 2000 total across dogs and humans) prior to *k*-nearest neighbors analysis and comparing the Spearman's correlation observed to that obtained for each permutation (**Supplementary Figure S3.2h-i**).

<u>Fitting the epigenetic age transfer function.</u> The nearest neighbor analysis was fit using non-linear regression with the SciPy package in Python. The model fit was specified using the following formula:

$$Dog\ Age = A * ln(Human\ age) + B$$

Here, “Dog age” was represented by the chronological ages of dogs, and “Human age” was the average age of the nearest human neighbors with respect to methylome similarity. The converse was performed as well, i.e. dog age was represented as the average age of the nearest dog neighbors and human age was the chronological age in humans. For the final age transfer function, the coefficients (A,B) were estimated by bootstrapping an equal number of both dogs and humans. The standard error was estimated using 1000 bootstraps.

Mouse validation of the conserved epigenetic progression. Methylome similarity was calculated identically, with dogs binned into two-year windows and mice binned into 0.5-year windows. A *k*-nearest neighbors analysis (as described for dogs and humans above) was repeated using the orthologous CpGs for pairwise comparisons involving mice. The mice represented in the mouse methylome data had a highly canalized age distribution which was different from that of the dogs or humans in our study. Multiple mice had been sampled at the exact same age, thereby discretizing mouse age into a small number of values. Mouse age was thus represented using five discrete bins.

Identification of gene orthologs with conserved methylation trajectories. We considered 14,652 one-to-one orthologs in dogs, humans and mice that were within 2.5kb of orthologous CpGs. Among these, we identified 7,934 orthologous genes for which methylation values were available. Methylation values were then logit-transformed; multiple CpGs assigned to one gene were represented by the average methylation value. We assigned to each ortholog a conservation score using the following procedure. First, the age of each dog and mouse individual was translated to the equivalent human age using the epigenetic age translation functions built using the *k*-nearest neighbors analysis. We then ranked all individuals according to their age in human years and divided this ranking into 15 quantile bins. Logit-transformed methylation values were averaged within each bin and species. For each gene and species we calculated the Spearman's correlation between the gene's methylation values and age. Genes were then ranked by $sign(correlation) * -log_{10}(correlation_{p\ value})$within each of the three species. We computed the Euclidean norm of the three ranks and sought genes with very low norms (for which methylation was consistently among the most increasing with age across species) or with very high norms (for which methylation was consistently among the most decreasing with age across species). Significance was determined using a two-sided empirical p-value < 0.05, yielding 394 genes.

Network analysis. We downloaded a composite human functional interaction network from (Huang et al. 2018), and sub-selected the network to only include significant CABs resulting in 355 nodes and 2003 edges. We visualized the network using Cytoscape 3.7 and performed community detection using clusterMaker2 (Shannon et al. 2003). To annotate modules, we performed functional enrichment using a hypergeometric test for each term within the Biological Process branch of the human Gene Ontology (GO) (Ashburner et al. 2000) and adjusted for false-discovery rate using Benjamini-Hochberg (FDR < 0.001). These results were clustered according to gene-set similarity using Enrichment Map (Merico, Isserlin, and Bader 2011), and modules were clustered according to the Jaccard overlap, revealing high-level functional categories (**Figure 3.3**).

Developmental genes analysis. For analysis of sequence constraint, We then ranked genes according to their aging conservation score and subdivided these genes into 25 evenly spaced bins, separating genes that were identified as significantly conserved, for a total of 27 bins. We the obtained PhyloP(Siepel, Pollard, and Haussler 2006) scores (described above) and extracted the PhyloP score according to the orthologous CpGs assigned to each orthologous gene. Finally, we averaged the phyloP scores according to their developmental gene status and their aging conservation score bin, estimating the 95% confidence interval by bootstrapping (**Supplementary**

Figure 3.5). We assessed the significance of the interaction between conservation score bin and developmental gene status using ANOVA.

For analysis assessing the effect of developmental genes on the epigenetic progression, we restricted to orthologous CpGs profiled across dogs, humans and mice (6,906 CpGs) that were within 2.5kb of the gene bodies of all orthologous genes (‘all CpGs’). From this set, we identified CpGs near development genes (‘devCpGs’); we also controlled for the number of CpGs with 100 randomly-sampled subsets of CpGs that were equal in size from those not near developmental genes (‘not devCpGs’). We calculated the methylome similarity (as described above) based on these CpG subsets for all three pairwise comparisons of species (dog and human, human and mouse, mouse and human). For each pairwise comparison (Species 1, Species 2), we identified the 5-nearest neighbors in Species 2 for each individual of Species 1, then binned the actual age of Species 1 into five discrete bins and calculated the average neighbor age for each bin with the 95% confidence interval estimated by bootstrapping.

Conserved Clock analysis. We built epigenetic clocks to measure dog years with Elastic net (scikit-learn in Python) where 85 dogs and 439 CpGs, which were assigned to CABs and profiled across three species, were used for training. For comparison, we randomly sampled 439 CpGs that were profiled across three species for a total of 100 randomly selected CpGs. We selected hyperparameters using 5-fold cross validation in the dog dataset. We assessed performance of the final model by using the Spearman correlation between actual age and the methylation age of 11 dogs, which were not used for training, and in the same mice that were previously described. We refer to the ages predicted from this model as “epigenetic ages”.

For analysis involving long-lived mice, we obtained DNA methylation data profiled from whole blood from (Petkovich et al. 2017) with the following counts described in **Supplementary**

Table 3.2 and processed identically as described above. We assessed significance of longevity-promoting interventions with a log-likelihood ratio test for all long-lived mice and their age matched controls. To determine strain or condition-specific effects, we calculated a control strain-specific average epigenetic age according to 10 evenly-spaced age bins. The epigenetic age of all mice assigned to the same bin were divided by this average, thereby reflecting the epigenetic relative aging ratio. Significance was assessed using a one-sided t-test.

3.6 Figures

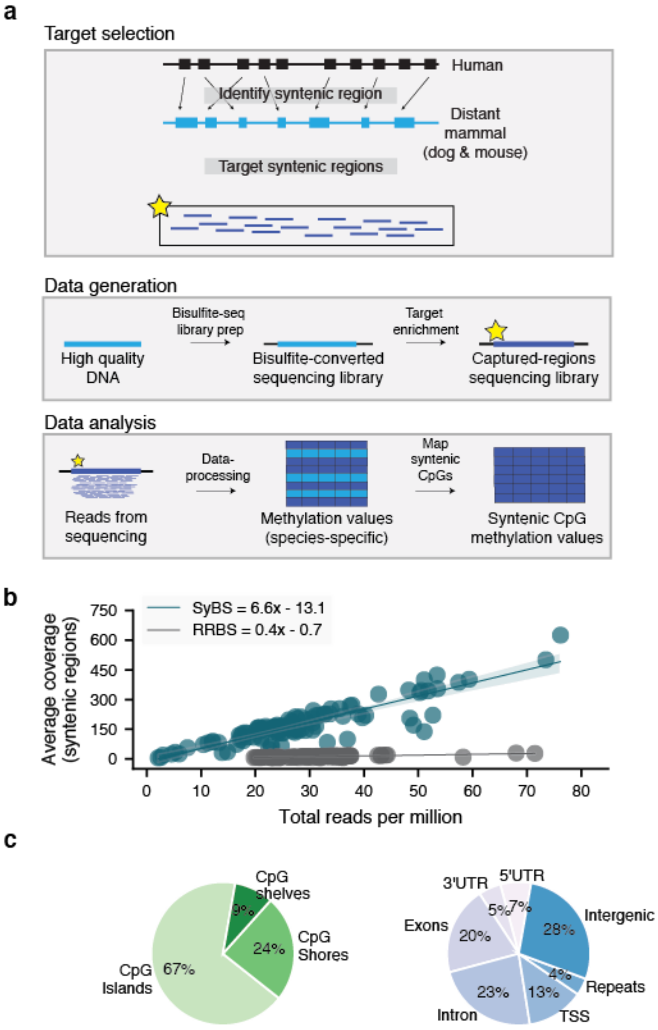


Figure 3.1: Profiling of the aging canine methylome by syntenic bisulfite sequencing (SyBS). (**a**) Strategy used to profile and compare CpG methylation within mammalian syntenic blocks. (**b**) Average coverage of syntenic segments versus total reads in millions, contrasting SyBS with Reduced Representation Bisulfite Sequencing (RRBS). (**c**) Pie charts showing representation of targeted genomic regions.

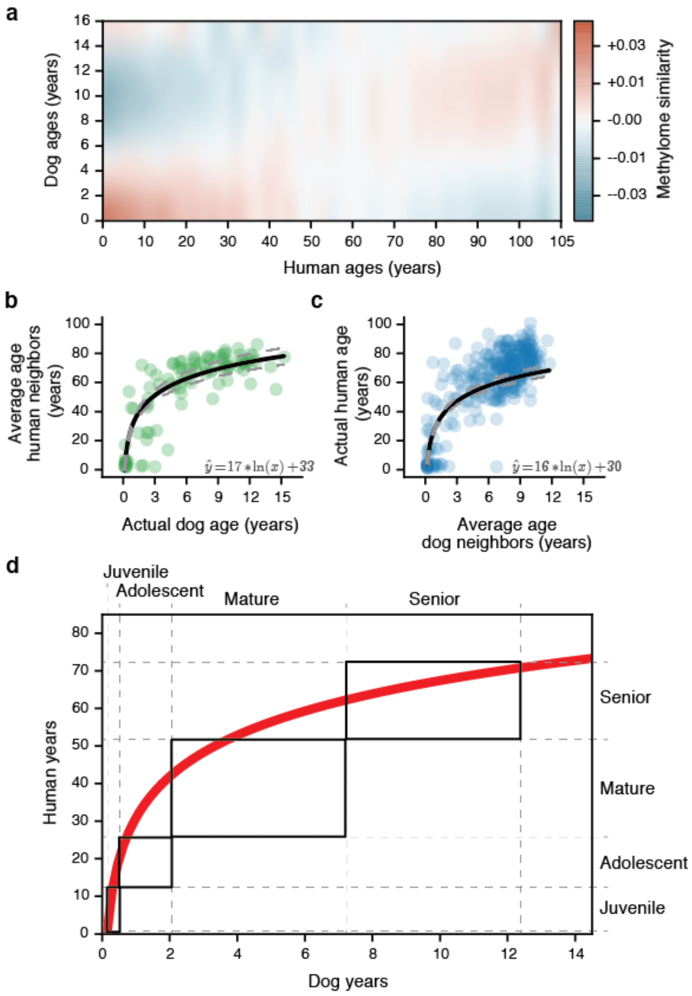


Figure 3.2: A non-linear transformation from dog to human age.
(**a**) Dog-human methylome similarities (Pearson's correlation) with dogs and humans averaged according to two-year windows. The average methylome correlation is shown for each cross-species pairing. Note that these correlations are over the entire methylome in an unsupervised analysis, revealing a low but significant methylome-wide similarity with age. (**b**) The age of each dog methylome (x-axis) is plotted against the average age of the five nearest human methylomes (y-axis). (**c**) Reciprocal plot in which the age of each human methylome (y-axis) is plotted against the average age of the five nearest dog methylomes (x-axis). (**d**) Logarithmic function for molecular translation from dog years (x-axis) to human years (y-axis). Outlined boxes indicate the approximate age ranges of documented life stages corresponding to common aging physiology. Juvenile refers to the period after infancy and before puberty, 2-6 mos. in dogs, 1-12 yrs. in humans; Adolescent refers to the period from puberty to completion of growth, 6 mos. to 2 yrs. in dogs, approximately 12-25 yrs. in humans; Mature refers to the period from 2-7 yrs. in dogs and 25-50 yrs. in humans; Senior refers to the subsequent period until life expectancy, 12 yrs. in dogs, 70 yrs. in humans. Dog stages are based on veterinary guides and mortality data for dogs (Bartges et al. 2012; Inoue et al. 2015; Fleming, Creevy, and Promislow 2011). Human stages are based on literature summarizing human life cycle and lifetime expectancy (Bogin and Smith 1996; Cia 2013; Arias, Heron, and Xu 2017) .

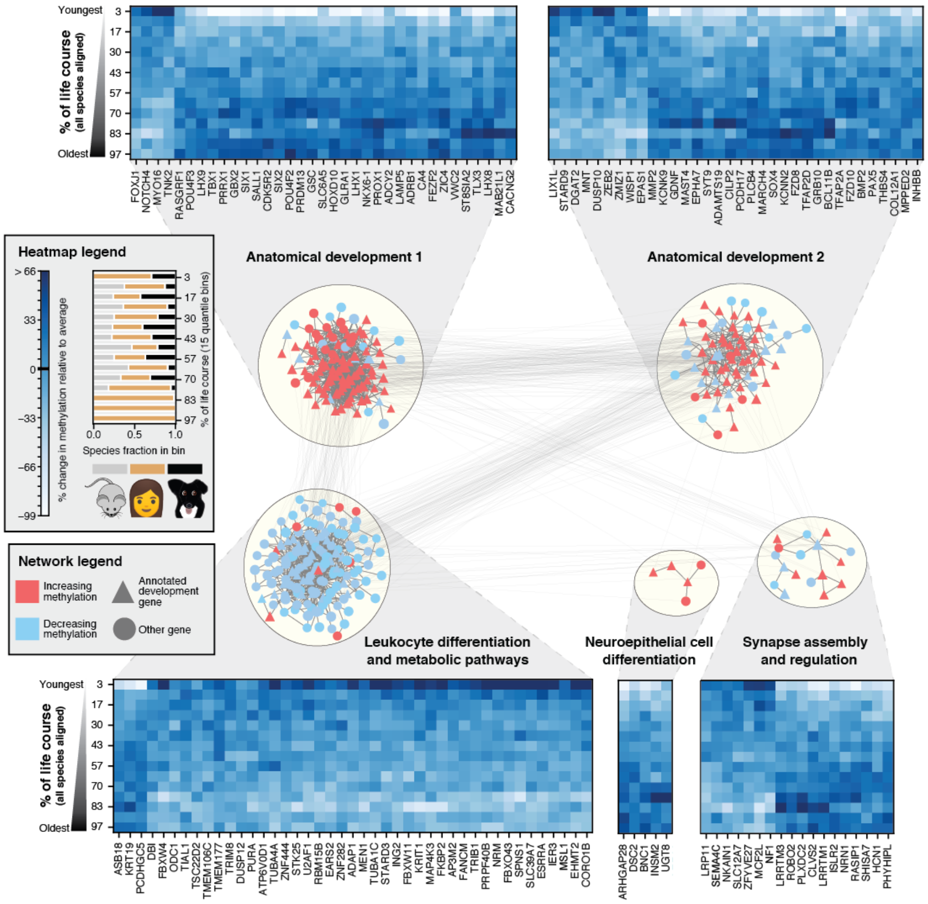


Figure 3.3: Conserved methylation changes clusters with one another in a functional interaction network.

The 394 genes exhibiting conserved age-related methylation behavior were mapped onto a human composite network in which edges represent functional interactions supported by multiple sources, resulting in 290 genes and 2003 edges. The network was clustered using community detection and enrichment in biological pathways (Gene Ontology, see **Methods**), resulting in 5 major modules that are labeled according to the enrichment. The colors represent the conserved direction of change with age, with red representing genes that increase methylation with age, and blue are those that decrease methylation with age. The heatmap underneath each module show the conserved methylation patterns of a randomly sampled subset of the module. Rows represent orthologs, columns represent the average value of all species ranked according to their age in human years divided into 15 age bins (quantiles) for all three species. Values are normalized according to the mean and standard deviation of methylation for each ortholog. The fractional species composition of each bin is visualized in the legend.

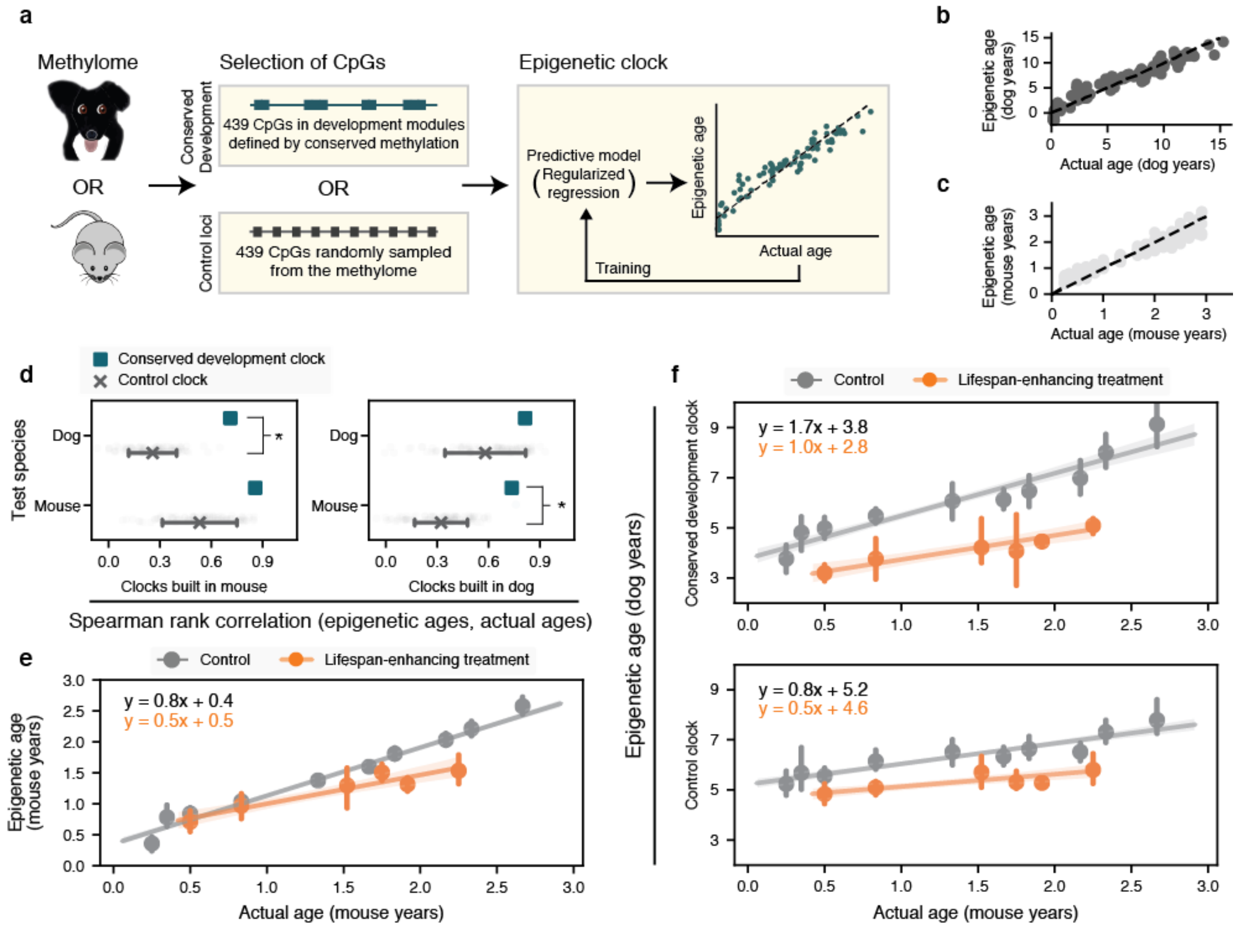
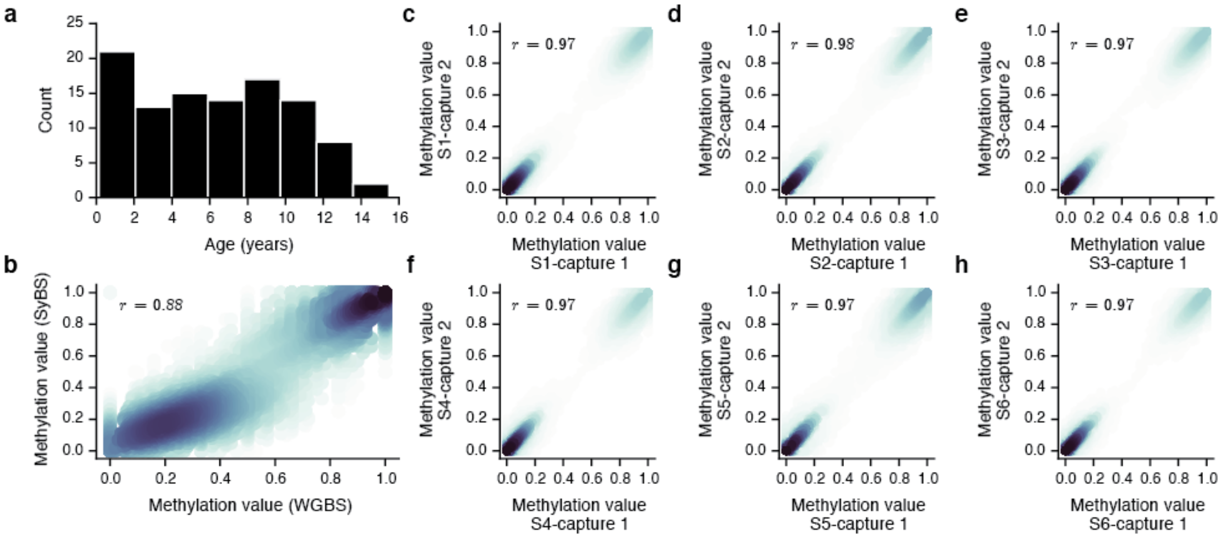


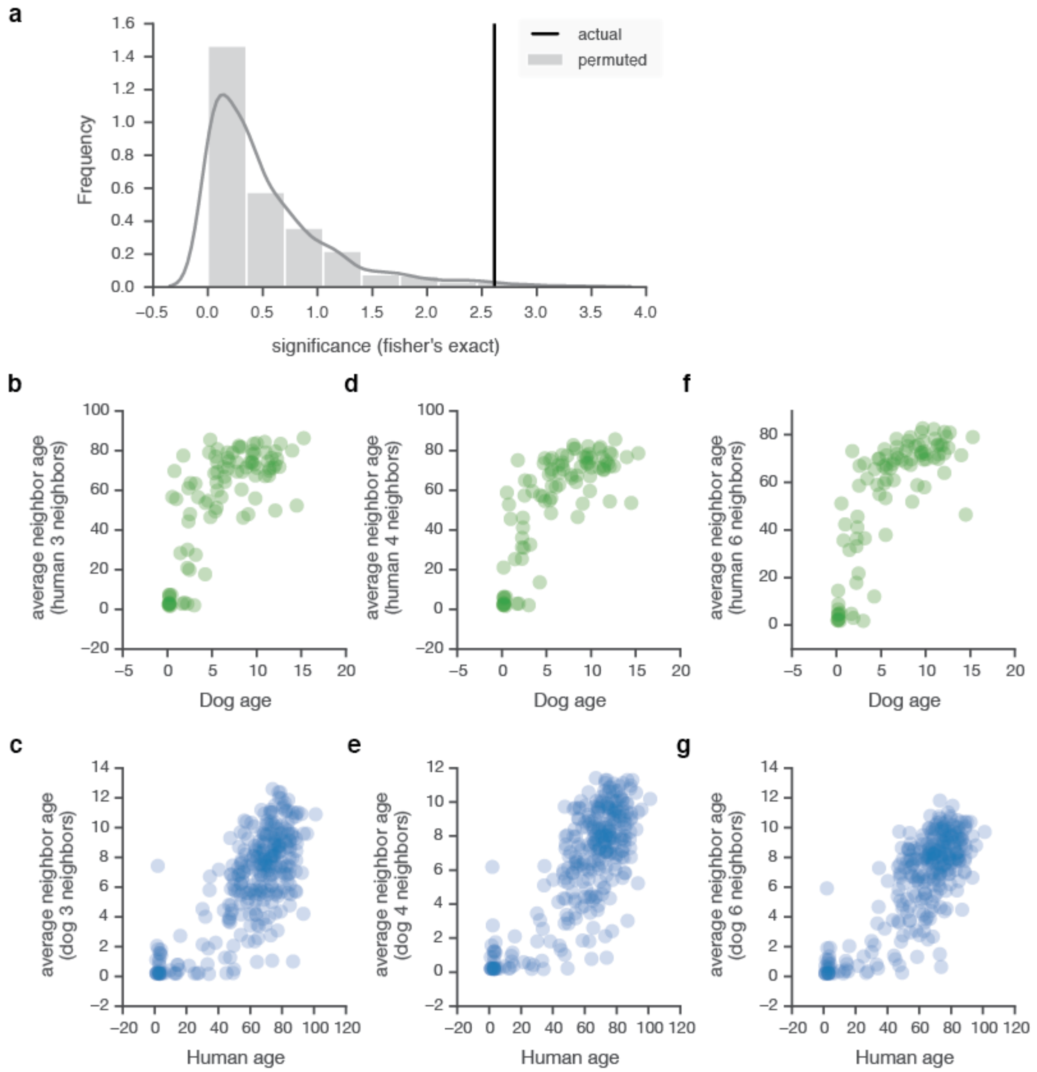
Figure 3.4: Conserved clocks measures biological aging effects.

(**a-c**) Schematic illustrating the epigenetic clock construction method. The method takes as input, the methylation values from either mice or dogs, using the methylation values from CpGs under genes exhibiting conserved methylation changes with age (Conserved development) or by randomly sampling the same number of CpG sites (Control clock). This is then trained using a regularized regression framework (Elastic Net) using the methylation values to predict the chronological age of the training species. The fully trained model is referred to as an " epigenetic clock" that measures the age of the training species, in either (**b**) dog years or (**c**) mouse years. (**d**) The Spearman's correlation between the epigenetic age and actual age when training clocks in mice or in dogs for the test species indicated on the y-axis. Different colors represent the different subsets of CpGs, either conserved clocks or random clocks according to the colors in (**a**). The average correlation of 100 randomly sampled subsets for random clocks is shown and the 95% confidence interval estimated for bootstrapping. (**e**) The conserved clock trained in mice and applied to long-lived mouse data, where the mice are binned according to 10 quantile bins, where the actual age and epigenetic age are averaged according to these bins and their longevity-promoting treatments. The best fit line, corresponding to the equations, of control mice or mice treated with longevity-interventions are shown, with colors reflected in the legend. The bands and bars depict the 95% confidence interval. (**f**) The actual age of mice and their epigenetic ages measured in dog years when using the conserved clock (top) or the random sampled subsets of CpGs (bottom). The same representation is depicted as (**e**). * denotes $p < 0.05$.

3.7 Supplementary Figures and Tables

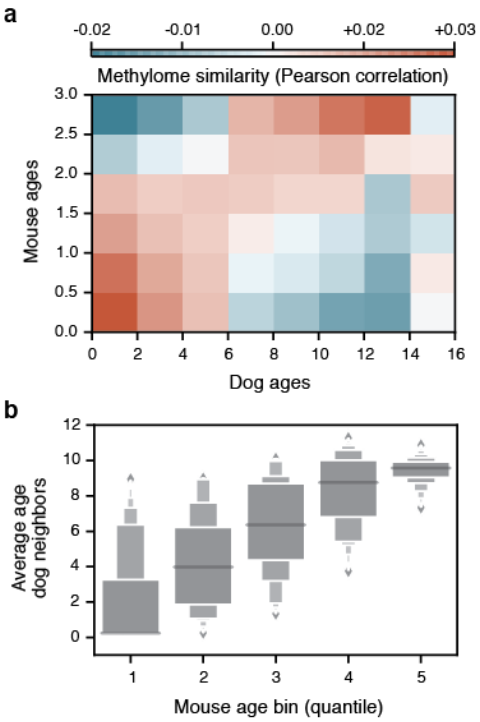


Supplementary Figure 3.1: Concordance of SyBS with non-targeted sequencing.
(a) The distribution of ages of 104 dogs used in this study is depicted as a histogram: the x-axis shows the age in years, the y-axis describes the number for each bin. **(b)** 10 samples were sequenced without enrichment for syntenic regions (SyBS), representing methylation values obtained from whole-genome bisulfite sequencing (WGBS). The methylation values for each sample before (x-axis) versus after (y-axis) hybridization are shown as a scatterplot where color represents the density of observations at each point (darker colors represent higher densities). Sites were considered if they were covered by >5 reads for both SyBS and WGBS. (**c-h**) Concordance of SyBS values for six canine DNA samples (S1-S6), for which two independent captures were performed. The x-axis represents values obtained for the first capture, and the y-axis represents values obtained for the second capture. The Pearson's correlation of the two captures is shown. Colors represent density of observations as per panel **b**.



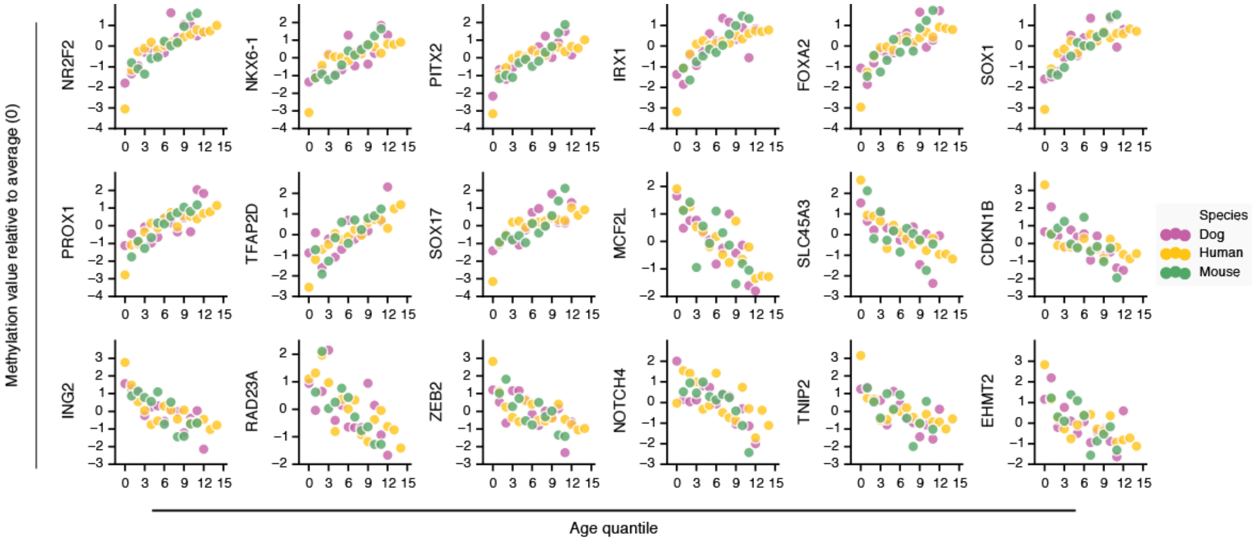
Supplementary Figure S3.2: Evaluating methylome similarities observed for dogs and humans.

(**a**) Black line: the observed p-value of association between methylome similarity for each pair of species and age. P-value computed by Fisher's exact test. This p-value is compared to those obtained from 1000 randomizations in which dog and human two-year bins were permuted (gray bars). (**b-g**) Varying the number of nearest neighbors *k* in dog-to-human age alignments. The actual ages of dogs (in years, x-axis, **a,c,e**) or humans (x-axis, **b,d,f**) versus the average age of *k* human or dog neighbors (in years, y-axis), respectively, when varying *k*: (**a,b**) *k* = 3, (**c,d**) *k* = 4 and (**e,f**) *k* = 6.



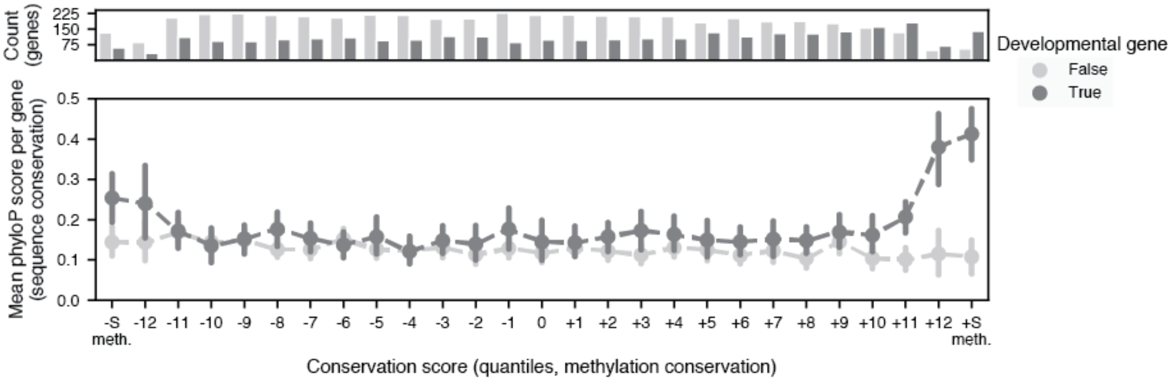
Supplementary Figure S3.3: Mouse methylation is concordant with conserved progressions observed in dogs and humans.

(**a**) Methylome similarity for each dog-mouse pair (Pearson's correlation) averaged according to 2-year bins for dogs (x-axis) and 0.5-year bins for mouse (y-axis). (**b**) Data in panel **a** are summarized by sorting mice into 5 age quantiles (x-axis) and, for each quantile, providing the distribution of the average age of the 5 nearest dogs according to methylome similarity. These values are shown in a box plot, where the largest box represents the 25th percentile to 75th percentile. The boxes outside of these values are quantiles scaled to the proportion of observations within each box. The horizontal line represents the median.



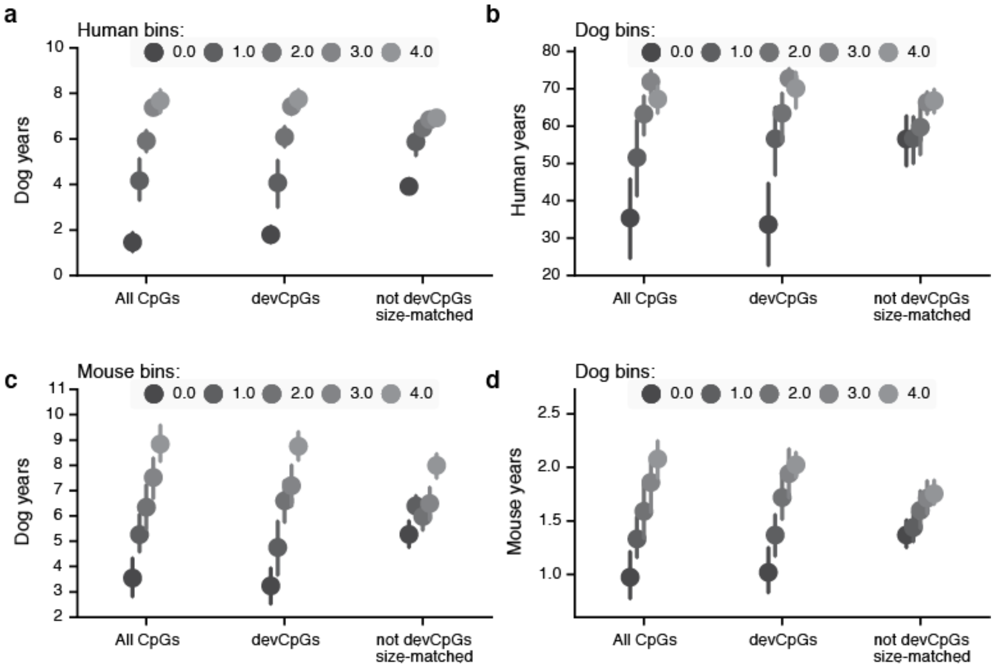
Supplementary Figure S3.4: Examples of genes exhibiting conserved methylation changes with age.

18 out of 394 orthologs exhibiting conserved direction of change with age are shown. All individuals are ranked according to their age in dog-years, then sorted and assigned into 15 age bins (quantiles). The x-axis shows these age bins. The y-axis shows the methylation change relative to average (0) where all values have been normalized by the mean and standard deviation within each species. The colors indicate the species.



Supplementary Figure S3.5: Relationship between sequence constraint and conserved methylation changes with age for developmental genes.

The x-axis depicts the aging conservation score for 7,942 genes, which are divided into 25 quantile bins. Those that are significant are specified by -S. meth, for significantly decreasing methylation with age and +S. meth are those that are significantly increasing methylation with age, consisting of 198 and 196 genes, respectively. The y-axis depicts the average PhyloP score for all genes within a bin, stratified according to whether the gene is a developmental gene. The bars indicate the 95% confidence interval estimated by bootstrapping. The bars depict the number of genes within each bin, colored according to their developmental status.



Supplementary Figure S3.6: Dependence of conserved epigenetic progression with CpGs within developmental modules.

The plots show epigenetic life trajectories based on methylome similarities calculated using all CpGs near genes (All CpGs), CpGs near development genes (devCpGs), or randomly sampled sets of CpGs not near development genes, of equal size to devCpGs (not devCpGs size-matched). The panels correspond to the following species comparisons: humans versus dogs (**a**), dogs versus humans (**b**), mice versus dogs (**c**), dogs versus mice (**d**), mice versus humans (**e**), and humans versus mice (**f**). In each comparison, the x-axis represents quintile age bins for one species, and the y-axis represents the average neighbor age for all individuals in each bin, where neighbor age is determined by the average age of the 5-nearest neighbors. Vertical bars represent the 95% confidence intervals of the averages obtained from 1000 bootstrapped samplings.

Supplementary File S3.1: Dogs used in study.

The file lists the 104 dogs used in this study: the sample ID, age in years and center where the sample was obtained are shown. Abbreviations: Labrador retriever (Lab), Mixed breed (Mix), UC Davis (Davis), National Human Genome Research Institute (NHGRI), Consented Volunteer (CV).

Supplementary File S3.2: Mouse dataset description.

This file describes the chronological age and longevity intervention for all mice that were compared in Figure 3.4.

3.8 Author Contributions

TW and TI initiated and conceptualized the study. TW carried out all experiments and implemented the main analyses. AH, EAO and DB provided canine samples. JM, SF, BT, JFK, DB and ARC assisted with miscellaneous analyses and provided feedback. PDA provided key input on human and animal aging. TW, TI, PDA, and EAO interpreted results and wrote the manuscript.

3.9 Acknowledgements

We gratefully acknowledge support for this work by grants from the California Institute for Regenerative Medicine (CIRM Epigenome Center to TI), the National Institute for Environmental Health Sciences (TI) and the Intramural Program of the National Human Genome Research Institute (AH and EAO). Competing Interests: TI is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. TI is on the Scientific Advisory Board of Ideaya BioSciences, Inc., has an equity interest, and receives sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies. Data Availability: All data and codes for analysis are available in Supplementary Materials and/or by request from the authors. Data materials availability: Sequencing data for dogs will be deposited into GEO and the accession number will be added prior to publication.

Chapter 3, in full, is currently being prepared for submission of the material and may appear as “A conserved epigenetic progression aligns dog and human age” by Tina Wang, Jianzhu Ma, Andrew Hogan, Samson Fong, Brian Tsui, Jason F. Kreisberg, Peter D. Adams, Anne-Ruxandra

Carvunis, Danika Bannasch, Elaine A. Ostrander and Trey Ideker. The dissertation author was the primary investigator and author of this material.

3.10 References

Alisch, Reid S., Benjamin G. Barwick, Pankaj Chopra, Leila K. Myrick, Glen A. Satten, Karen N. Conneely, and Stephen T. Warren. 2012. “Age-Associated DNA Methylation in Pediatric Populations.” *Genome Research* 22 (4): 623–32.

Andrews, Simon, and Others. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.”

Arias, Elizabeth, Melonie Heron, and Jiaquan Xu. 2017. “United States Life Tables, 2013.” *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 66 (3): 1–64.

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. “Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays.” *Bioinformatics* 30 (10): 1363–69.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29.

“Babraham Bioinformatics - Trim Galore!” 2016. Accessed December 17. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Bartges, Joe, Beth Boynton, Amy Hoyumpa Vogt, Eliza Krauter, Ken Lambrecht, Ron Svec, and Steve Thompson. 2012. “AAHA Canine Life Stage Guidelines.” *Journal of the American Animal Hospital Association* 48 (1): 1–11.

Bell, G. I., J. H. Karam, and W. J. Rutter. 1981. “Polymorphic DNA Region Adjacent to the 5’ End of the Human Insulin Gene.” *Proceedings of the National Academy of Sciences of the United States of America* 78 (9): 5759–63.

Bogin, Barry, and B. Holly Smith. 1996. “Evolution of the Human Life Cycle.” *American Journal of Human Biology: The Official Journal of the Human Biology Council* 8 (6): 703–16.

Bolstad, Benjamin Milo. 2013. “preprocessCore: A Collection of Pre-Processing Functions.” *R Package Version* 1 (0).

Cia, U. 2013. “The World Factbook 2013-14.” *Central Intelligence Agency*.

Cohen, Alan A. 2018. "Aging across the Tree of Life: The Importance of a Comparative Perspective for the Use of Animal Models in Aging." *Biochimica et Biophysica Acta, Molecular Basis of Disease* 1864 (9 Pt A): 2680–89.

Davis, Brian W., and Elaine A. Ostrander. 2014. "Domestic Dogs and Cancer Research: A Breed-Based Genomics Approach." *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources* 55 (1): 59–68.

Dedeurwaerder, Sarah, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou, and François Fuks. 2011. "Evaluation of the Infinium Methylation 450K Technology." *Epigenomics* 3 (6): 771–84.

Dreger, Dayna L., Maud Rimbault, Brian W. Davis, Adrienne Bhatnagar, Heidi G. Parker, and Elaine A. Ostrander. 2016. "Whole-Genome Sequence, SNP Chips and Pedigree Structure: Building Demographic Profiles in Domestic Dog Breeds to Optimize Genetic-Trait Mapping." *Disease Models & Mechanisms* 9 (12): 1445–60.

"Fancyimpute." n.d. https://github.com/iskandr/fancyimpute.

Field, Adam E., Neil A. Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D. Adams. 2018. "DNA Methylation Clocks in Aging: Categories, Causes, and Consequences." *Molecular Cell* 71 (6): 882–95.

Fleming, J. M., K. E. Creevy, and D. E. L. Promislow. 2011. "Mortality in North American Dogs from 1984 to 2004: An Investigation into Age-, Size-, and Breed-Related Causes of Death." *Journal of Veterinary Internal Medicine / American College of Veterinary Internal Medicine* 25 (2): 187–98.

Gilmore, Keiva M., and Kimberly A. Greer. 2015. "Why Is the Dog an Ideal Model for Aging Research?" *Experimental Gerontology* 71 (November): 14–20.

Gross, A. M., P. A. Jaeger, J. F. Kreisberg, K. Licon, K. L. Jepsen, M. Khosroheidari, B. M. Morsey, S. Swindells, H. Shen, C. T. Ng, K. Flagg, D. Chen, K. Zhang, H. S. Fox, and T. Ideker. 2016. "Methylome-Wide Analysis of Chronic HIV Infection Reveals Five-Year Increase in Biological Age and Epigenetic Targeting of HLA." *Molecular Cell* 62 (2): 157–68.

Hannum, G., J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J. B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang. 2013. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates." *Molecular Cell* 49 (2): 359–67.

Hastie, Trevor, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. 2011. "Impute: Impute: Imputation for Microarray Data." R package version.

Horvath, S. 2013. "DNA Methylation Age of Human Tissues and Cell Types." *Genome Biology* 14 (10): R115.

Huang, Justin K., Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, and Trey Ideker. 2018. "Systematic Evaluation of Molecular Networks for

Discovery of Disease Genes." *Cell Systems* 6 (4): 484–95.e5.

Inoue, Mai, A. Hasegawa, Y. Hosoi, and K. Sugiura. 2015. "A Current Life Table and Causes of Death for Insured Dogs in Japan." *Preventive Veterinary Medicine* 120 (2): 210–18.

Jones, Eric, Travis Oliphant, Pearu Peterson, and Others. 2015. "SciPy: Open Source Scientific Tools for Python, 2001." *URL Http://www. Scipy. Org* 73: 86.

Kaeberlein, Matt, Kate E. Creevy, and Daniel E. L. Promislow. 2016. "The Dog Aging Project: Translational Geroscience in Companion Animals." *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 27 (7-8): 279–88.

Kirkwood, Thomas B. L. 2005. "Understanding the Odd Science of Aging." *Cell* 120 (4): 437–47.

Krueger, Felix, and Simon R. Andrews. 2011. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." *Bioinformatics* 27 (11): 1571–72.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25.

Lebeau, A. 1953. "L'âge Du Chien et Celui de L'homme. Essai de Statistique Sur La Mortalité Canine." *Bulletin de l'Academie Veterinaire de France* 26: 229–32.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Lopez-Otin, C., M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. 2013. "The Hallmarks of Aging." *Cell* 153 (6): 1194–1217.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A. Baccarelli, Yun Li, James D. Stewart, Eric A. Whitsel, Themistocles L. Assimes, Luigi Ferrucci, and Steve Horvath. 2019. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan." *Aging* 11 (2): 303–27.

Madar, Aviv, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. 2010. "DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator." *PloS One* 5 (3): e9803.

Merico, Daniele, Ruth Isserlin, and Gary D. Bader. 2011. "Visualizing Gene-Set Enrichment Results Using the Cytoscape Plug-in Enrichment Map." *Methods in Molecular Biology*. doi:10.1007/978-1-61779-276-2_12.

"MethylDackel." n.d. https://github.com/dpryan79/MethylDackel.

Ostrander, Elaine A., Robert K. Wayne, Adam H. Freedman, and Brian W. Davis. 2017. "Demographic History, Selection and Functional Diversity of the Canine Genome." *Nature Reviews. Genetics* 18 (12): 705–20.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (November). JMLR.org: 2825–30.

Petkovich, Daniel A., Dmitriy I. Podolskiy, Alexei V. Lobanov, Sang-Goo Lee, Richard A. Miller, and Vadim N. Gladyshev. 2017. "Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions." *Cell Metabolism* 25 (4): 954–60.e6.

"Picard Tools - By Broad Institute." 2017. Accessed March 15. http://broadinstitute.github.io/picard.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

Rosenbloom, Kate R., Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Glenn Hickey, Angie S. Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T. Lee, Chin H. Li, Karen H. Miga, Ngan Nguyen, Benedict Paten, Brian J. Raney, Arian F. A. Smit, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. 2015. "The UCSC Genome Browser Database: 2015 Update." *Nucleic Acids Research* 43 (Database issue): D670–81.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.

Siepel, Adam, Katherine S. Pollard, and David Haussler. 2006. "New Methods for Detecting Lineage-Specific Selection." In *Research in Computational Molecular Biology*, 190–205. Springer Berlin Heidelberg.

Skene, Nathan G., Marcia Roy, and Seth Gn Grant. 2017. "A Genomic Lifespan Program That Reorganises the Young Adult Brain Is Targeted in Schizophrenia." *eLife* 6 (September). doi:10.7554/eLife.17915.

Stubbs, Thomas M., Marc Jan Bonder, Anne-Katrien Stark, Felix Krueger, BI Ageing Clock Team, Ferdinand von Meyenn, Oliver Stegle, and Wolf Reik. 2017. "Multi-Tissue DNA Methylation Age Predictor in Mouse." *Genome Biology* 18 (1): 68.

Teschendorff, Andrew E., Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. "A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 K DNA Methylation Data." *Bioinformatics* 29 (2): 189–96.

Thompson, Michael J., Bridgett vonHoldt, Steve Horvath, and Matteo Pellegrini. 2017. "An Epigenetic Aging Clock for Dogs and Wolves." *Aging* 9 (3): 1055–68.

Vidaki, Athina, David Ballard, Anastasia Aliferi, Thomas H. Miller, Leon P. Barron, and Denise Syndercombe Court. 2017. "DNA Methylation-Based Forensic Age Prediction Using Artificial Neural Networks and next Generation Sequencing." *Forensic Science International. Genetics* 28 (May): 225–36.

Vilella, Albert J., Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. 2009. "EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates." *Genome Research* 19 (2): 327–35.

Vonholdt, Bridgett M., John P. Pollinger, Kirk E. Lohmueller, Eunjung Han, Heidi G. Parker, Pascale Quignon, Jeremiah D. Degenhardt, Adam R. Boyko, Dent A. Earl, Adam Auton, Andy Reynolds, Kasia Bryc, Abra Brisbin, James C. Knowles, Dana S. Mosher, Tyrone C. Spady, Abdel Elkahloun, Eli Geffen, Malgorzata Pilot, Wlodzimierz Jedrzejewski, Claudia Greco, Ettore Randi, Danika Bannasch, Alan Wilton, Jeremy Shearman, Marco Musiani, Michelle Cargill, Paul G. Jones, Zuwei Qian, Wei Huang, Zhao-Li Ding, Ya-Ping Zhang, Carlos D. Bustamante, Elaine A. Ostrander, John Novembre, and Robert K. Wayne. 2010. "Genome-Wide SNP and Haplotype Analyses Reveal a Rich History Underlying Dog Domestication." *Nature* 464 (7290): 898–902.

Wang, Meng, and Bernardo Lemos. 2019. "Ribosomal DNA Harbors an Evolutionarily Conserved Clock of Biological Aging." *Genome Research* 29 (3): 325–33.

Wang, Tina, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Michael Ku Yu, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. 2017. "Epigenetic Aging Signatures in Mice Livers Are Slowed by Dwarfism, Calorie Restriction and Rapamycin Treatment." *Genome Biology* 18 (1): 57.

Yates, Andrew, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. 2016. "Ensembl 2016." *Nucleic Acids Research* 44 (D1): D710–16.

CHAPTER 4: Discussion

4.1 Summary

This work has contributed to our understanding of the dynamics of epigenetic evolution. Since this work is mostly descriptive, it is still unclear exactly how these changes precisely affect cellular function and organism fitness. Nevertheless, this work has contributed towards delineating broad trends that are observed across many species.

In the first chapter, we harnessed naturally occurring differences in genomic architecture between *Mammalia* and *Drosophila* to understand the relationship ask whether the evolution of epigenetic features also followed the dynamics of sequence evolution (Villar, Flicek, and Odom 2014). Even though members of *Mus* have more rapid sequence evolution relative to members of *Drosophila,* we found that both clades had indistinguishable transcriptional evolution at different layers of gene regulatory networks (Carvunis et al. 2015). Notably, sequence-level features, including transcription factor binding and motifs, tended to diverge faster than the expression of genes. Nevertheless, this happened at indistinguishable rates in both clades. Our findings indicate that transcriptional evolution does not necessarily follow sequence evolution.

In the remaining chapters, I continued to study the evolution of epigenetic modifications, but at during the course of a lifetime. Previous studies had indicated that methylation could be used to measure age accurately in humans (Hannum et al. 2013; S. Horvath 2013). However, it was unclear whether this was a phenomena was only specific to humans or if these were applicable across all mammals. Therefore, in chapter 2, I asked whether this phenomena extended from humans to mice. For this purpose, I collapsed methylation data from three distinct studies (Reizel et al. 2015; Cannon et al. 2014; Gravina et al. 2016), where methylation was profiled in liver.

Using Elastic Net (Zou and Hastie 2005), a regularized regression framework, I built an accurate epigenetic clock where the predicted ages of mice were highly correlated with the actual age of the mice (Pearson's correlation = 0.91 in cross-validation). To verify that this measure recaptured aging-related effects, I applied this clock to mice treated with longevity-promoting interventions, specifically: rapamycin treatment, Ames dwarf mice and calorie restriction (Cole et al. 2017). In all mice, I saw that epigenetic ages measured by this clock were younger than those obtained for age matched controls. This work was the first demonstration that ages measured from these clocks could be slowed, indicating that some aspect of aging was captured through DNA methylation (Wang et al. 2017). Moreover, that epigenetic clocks were a conserved feature of the aging methylome across mammals.

After observing that conserved methylation changes changes that were conserved between mice and dogs, I then decided to build these clocks in dogs, and determine whether there were particular changes that were more conserved with respect to age. To be completely honest, I wanted to build molecular measures of age because I had just adopted my dog, and felt her age estimate was a bit suspicious. To my surprise, this actually launched the final two parts of my thesis. Nevertheless, this project has truly been a unique experience, consisting of the best and worst moments of my dissertation. The specific discovery I made was that genome-wide methylome similarities between individuals of dogs and humans could translate dog years to human years and vice versa. Motivated by these results, I then searched for the regions of the methylome that were responsible for these age-translations. I found that these conserved events predominantly affected development genes. Moreover, epigenetic clocks could be formulated when using only the CpGs under these conserved genes. At first, a simple explanation would that these genes were selected based on your criteria, effectively a circular pipeline. However, I validated this observation in long-lived mice and their controls, which were not used for identifying conserved methylation behavior. They also show the expected biological aging differences. Moreover, after further selecting using regularized regression, these CpGs were highly enriched for developmental genes. These findings indicate that precise changes in developmental genes are responsible for translating aging-related effects across species.

4.2 Limitations

There are important limitations to consider for this body of work. First and foremost, all of my work thus far has been descriptive. It is still unclear what the exact relationship between epigenetic changes are with cellular functions. This is still an important aspect that remains to be validated. In this section, I break down limitations by the studies with respect to the scale of time examined.

In chapter 1, the first and foremost limitation is that our conclusions are based off of a small sampling of individuals from each species in each clade. For instance, two individuals were sampled at each time point, meaning there was only 10 data points that could be fit by regression methods. It remains to be determined if larger population of individuals would change these findings. Moreover, since epigenetics varies with respect to cellular context, none of our stages were carefully matched. It remains to be determined if we would see different effects had we had access to data that was carefully matched with respect to life stages.

In chapter 2 and 3, beyond long-lived mice, we did not have any other metrics that would report other elements of ‘healthspan’, or the period of life when individuals are relatively healthy. Increasing the length of healthspan is the ultimate goal of understanding the mechanisms of aging. Finding reproducible molecular correlate of aging is incredibly exciting, but it is still in its infancy.

It still remains to be determined whether specific changes actually influence the rate of aging, and contribute towards the mechanisms of aging. Currently, methods to introduce methylation changes at defined loci are still in its infancy. However, there are studies that have been able to tether a de-methylating domain to CRISPR to introduce defined changes (Liu et al. 2016). These experiments will help define the functional effects of methylation changes and its contribution to aging.

Another important limitation to consider is that all of my findings, particularly in Chapter 3, were made using whole blood DNA isolates. It is well known that DNA methylation signatures have a strong dependency on changing cell types. It could be possible that conserved shifts in particular blood cell types are providing the signal to observe these changes across mammals. Nevertheless, it does make for an interesting model to understand how these conserved shifts are occurring.

4.3 Outlook

This section of my dissertation is fully dedicated to my personal opinions regarding the DNA methylation changes with age, and how this could revolutionize the ways we study aging.

Biology tends to define the period of development, prior to sexual maturity, and aging as two distinct periods of life, development and aging (Kirkwood 2005). Development is thought of as ‘programmed’, meaning that gene products are selected upon by natural selection to give rise to a sexually mature individual. In contrast, aging describes the progressive decline of physiological function and increased risk of mortality affecting nearly every species. Our understanding of aging has remained fundamentally limited, in part due to difficulties of defining aging, and the difficulties in understanding whether we should expect aging mechanisms to be conserved across species of extremely different lifespans.

These difficulties are fully driven by our inability to define aging in a way that is feasible to measure and study in practice. I believe that DNA methylation may provide the first molecular measure that can be used to report on the rate of aging, as we experience it. Having a measurement that correlates with elements of aging, such as lifetime-expectancy, can be immensely useful towards studying mechanisms that increase or decrease the rate of these changes. If this was true, this would reduce the burden of studying aging and truly understand the components that reduce our physiological fitness with respect to time. After all, having the ability to have a reproducible biomarker is necessary in order to further study any biological process.

If this is true, that DNA methylation somehow can track all the complex processes that break down over the span of life, it is natural to wonder why this is possible. I do not think that changes in DNA methylation with age are necessarily determined by a hypothetical aging-program that initiates during middle age. However, there may be some unintentional consequences of genetic programs that occur earlier in life, which result in aging. This theory of aging is formally defined as the *antagonistic pleiotropy of aging*, specifically that genes that were beneficial earlier in life become detrimental later in life (Steve Horvath and Raj 2018; Blagosklonny 2012). Consistent DNA methylation changes with age may eventually be evidence of this theory of aging. Perhaps, they represent a method by which 'genomes' have any sense of progression through the lifespan. These epigenetic changes may be defined by species-specific features that orchestrate the length of gestation, the length of time it takes to reach sexual maturation and the maximal longevity. Intriguingly, each of these traits are strongly and inversely associated with maximal lifespan, such that shorter-lived mammals typically have short gestation periods and reach sexual maturation at much faster rates. Moreover, the conserved and highly punctated increasing methylation with age, predominantly affecting development genes, are incredibly consistent, and much more consistent than those that decrease methylation with age, or the methylation around non-developmental genes. Perhaps, the amount of methylation at this loci somehow tells the genome how ‘old’ it is. It would be exciting to understand this phenomena and its relationship with aging. Wouldn’t it be fascinating if the reasons for why we age are directly related to our necessity to reach sexual maturation across all mammals? Moreover, what if one could determine what genetic elements control the rate in which this methylation gains with life? This might be all crazed musings, but I truly believe that this is an area that deserves further investigation to understand the ultimate basis for every other age-related disease.

4.4 References

Blagosklonny, Mikhail V. 2012. “Answering the Ultimate Question ‘what Is the Proximal Cause of Aging?’” *Aging* 4 (12). Impact Journals, LLC: 861.

Cannon, Matthew V., David A. Buchner, James Hester, Hadley Miller, Ephraim Sehayek, Joseph H. Nadeau, and David Serre. 2014. “Maternal Nutrition Induces Pervasive Gene Expression Changes but No Detectable DNA Methylation Differences in the Liver of Adult Offspring.” *PloS One* 9 (3): e90335.

Carvunis, Anne-Ruxandra, Tina Wang, Dylan Skola, Alice Yu, Jonathan Chen, Jason F. Kreisberg, and Trey Ideker. 2015. “Evidence for a Common Evolutionary Rate in Metazoan Transcriptional Networks.” *eLife* 4 (December). doi:10.7554/eLife.11615.

Cole, John J., Neil A. Robertson, Mohammed Iqbal Rather, John P. Thomson, Tony McBryan, Duncan Sproul, Tina Wang, et al. 2017. “Diverse Interventions That Extend Mouse Lifespan Suppress Shared Age-Associated Epigenetic Changes at Critical Gene Regulatory Regions.” *Genome Biology* 18 (1): 58.

Gravina, Silvia, Xiao Dong, Bo Yu, and Jan Vijg. 2016. “Single-Cell Genome-Wide Bisulfite Sequencing Uncovers Extensive Heterogeneity in the Mouse Liver Methylome.” *Genome Biology* 17 (1): 150.

Hannum, G., J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, et al. 2013. “Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.” *Molecular Cell* 49 (2): 359–67.

Horvath, S. 2013. “DNA Methylation Age of Human Tissues and Cell Types.” *Genome Biology* 14 (10): R115.

Horvath, Steve, and Kenneth Raj. 2018. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing." *Nature Reviews. Genetics* 19 (6): 371–84.

Kirkwood, Thomas B. L. 2005. "Understanding the Odd Science of Aging." *Cell* 120 (4): 437–47.

Liu, X. Shawn, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A. Young, and Rudolf Jaenisch. 2016. "Editing DNA Methylation in the Mammalian Genome." *Cell* 167 (1): 233–47.e17.

Reizel, Yitzhak, Adam Spiro, Ofra Sabag, Yael Skversky, Merav Hecht, Ilana Keshet, Benjamin P. Berman, and Howard Cedar. 2015. "Gender-Specific Postnatal Demethylation and Establishment of Epigenetic Memory." *Genes & Development* 29 (9): 923–33.

Villar, Diego, Paul Flicek, and Duncan T. Odom. 2014. "Evolution of Transcription Factor Binding in Metazoans - Mechanisms and Functional Implications." *Nature Reviews. Genetics* 15 (4): 221–33.

Wang, Tina, Brian Tsui, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Michael Ku Yu, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. 2017. "Epigenetic Aging Signatures in Mice Livers Are Slowed by Dwarfism, Calorie Restriction and Rapamycin Treatment." *Genome Biology* 18 (1): 57.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2). Blackwell Publishing Ltd: 301–20.