UNIVERSITY OF CALIFORNIA

Los Angeles

Prediction and Inference for High-Dimensional Genetic Data

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

by

Caesar Zexuan Li

2022

ABSTRACT OF THE DISSERTATION

Prediction and Inference

for High-Dimensional Genetic Data

by

Caesar Zexuan Li

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2022

Professor Gang Li, Chair

Collection of large amounts of genetic data and advancements in computational genetics over the recent years provide us with tools to explore epigenetic mechanisms that lead to aging and lifespan. In the context of continuous DNA methylation data, with a novel cross-species DNA methylation microarray targeting conserved CpG sites across mammalian species, we are able to leverage readily available statistical models to extensively study important life history traits such as lifespan, gestation time, and time to sexual maturity across various species. DNA methylation data are often high dimensional and require regularized regression frameworks to construct practical prediction models. Based on an unprecedented mammalian DNA methylation data set, we have developed methylation-based epigenetic life history traits predictors using regularized linear regressions. The estimators can accurately predict maximum lifespan using cytosine methylation patterns collected from over 13,000 samples derived from 348 mammalian species. To extend our future inferential

analyses into diverse data sources such as RNA-seq data, we have proposed an L_0-regularized Poisson graphical model for exploring gene-to-gene relations. The superior theoretical properties that the L_0 sparse graphical model enjoys will more effectively assist the future work of clustering and grouping large numbers of DNA methylation sites and genes. Both the applied research and methodological work will aid in the aging research goals of integrating various layers of multiomics data.

To my loving father, who passed on during my PhD study.

I wish I could travel with you one more time in an RV.

The dissertation of Caesar Zexuan Li is approved.

Jason Ernst

Zhe Fei

Donatello Telesca

Gang Li, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my doctoral advisors, Dr. Gang Li and Dr. Steve Horvath, for their guidance, support, and patience, during my academic career at UCLA. I have learned tremendously under their guidance. I am also extremely grateful to my current and former committee members: Drs. Zhe Fei, Jason Ernst, Donatello Telesca, and Janet Sinsheimer for their valuable comments and insights. I would like to extend my thanks to my family, my late father, Anqi Li, and my mother, Zhenzhu Guo, for their unconditional loving support. I couldn't have accomplished this without them. Last but not least, my special thanks go to my fiancée, Sydney Truong, who has always given me cheerful support for my career.

VITA

Education and employment

2017 – 2019 M.S. (Biostatistics), University of California, Los Angeles.

2012 – 2016 B.S. (Financial Actuarial Mathematics), University of California, Los Angeles.

2018 – 2022 Graduate Student Researcher, Department of Biostatistics, UCLA.

2018 – 2019 Teaching Assistant, Department of Biostatistics, UCLA.

Publications

Li, C. Z., Kawaguchi, E. S., & Li, G. (2021). A New L_0-Regularized Log-Linear Poisson Graphical Model with Applications to RNA Sequencing Data. *Journal of Computational Biology*.

Publications under review

Li, C., Haghani, A., Robeck, T. R., Villar, D., Lu, A. T., Zhang, J., … & Horvath, S. (2021). Epigenetic predictors of maximum lifespan and other life history traits in mammals. *bioRxiv*.

Haghani, A., **Li, C.**, Lu, A. T., Robeck, T. R., Belov, K., Breeze, C. E., ... & Horvath, S. (2021). DNA Methylation Networks Underlying Mammalian Traits. *bioRxiv*.

In preperation

Li, C., Haghani, A., Lu, A., Horvath, S. (2021). Alternative Approaches of Predictors and Epigenome-wide Association studies for Mammalian life history traits. *In preparation*.

CHAPTER 1

1. Introduction

In the age of rapidly growing technologies in both data collection and storage tools, the fields of statistics and genetics face many opportunities as well as challenges. Data sets with tens of thousands of samples and many times more variables are the new norm to statisticians in the field of genetics. Many classical statistical methods can be readily and robustly applied to these large data sets directly, such as genome-wide association studies (GWAS) for single-nucleotide polymorphisms (SNP) array data, and epigenome-wide association studies (EWAS) for DNA methylation data. The idea is to evaluate a phenotypical trait against genotypes one locus at a time, and then summarize findings after adjusting for multiple hypothesis testing using, most commonly, false discovery rates (FDR) or Bonferroni correction (Y. Benjamini, and Hochberg, Y., 1995; C. E. Bonferroni, 1935). These methods have been effective in discovering associations between individual genotypes and phenotypes, leading to numerous discoveries and remarks (Uffelmann et al., 2021; Visscher et al., 2017). Some improvements on GWAS, taking environmental factors into account, include Newton's method and scoring method used for polygenic models, which are enabled by linear mixed models (Lange, 2003). For multi-marker analyses in single nucleotide polymorphisms (SNPs) data, researchers have aggregated the effects across all loci using a modified linear mixed model to include all SNPs simultaneously (J. Yang et al., 2010), or directly summing over GWAS effect size results to form a polygenic risk score (PRS) (Dudbridge, 2013; Palla & Dudbridge, 2015). For epigenetics, some of the more data-driven multivariate models used in recent years include elastic net for DNA methylation data (S. Horvath, 2013; Ake T Lu et al., 2019; Zou & Hastie, 2005). This dissertation focuses on DNA methylation and RNA-seq data

analyses. In Chapters 3 and 4, we describe the applied research that has been done to DNA methylation data in the field of aging research, with goals of studying mammalian species life history traits, such as species maximum lifespan, gestation time, and time to sexual maturity. In Chapter 5, we introduce a novel methodology in the form of a Poisson graphical model for RNA-seq data. In Chapter 6, we address both the current and new statistical methods and outline possible future work that will compare and potentially improve these algorithms.

1.1. DNA methylation data

DNA methylation is an epigenetic mechanism by which methyl groups are attached to DNA molecule. This process can regulate gene expressions from a DNA segment without changing the sequence. This is achieved by either preventing transcription factors' binding to the sequence or recruiting proteins that are involved in gene expression (Moore, Le, & Fan, 2013). Most of DNA methylation occur on cytosine nucleotide that immediately precedes a guanine. The sites at which DNA methylation occur are called cytosine-phosphate-guanine (CpG) sites. Researchers collect DNA methylation data by using a methylation array that has thousands of probes designed to detect the intensity of each site's methylation. For example, one of the most efficient and comprehensive methylation array for human genome, Illumina 450k, has over 480k CpG probes, providing practically whole-genome coverage (Bibikova et al., 2011).

Raw DNA methylation data are collected as florescent intensity measurements from methylated and unmethylated probes. After background adjustment and normalization, both methylated and unmethylated intensity measurements at each site are converted to a single value, either a beta-value or an m-value. The more common method is the beta-value, which is the ratio

of the methylated intensity to the sum of methylated, unmethylated intensities, and a constant of 100 as an offset for stabilizing sites where both intensities are small (Du et al., 2010),

(Equation 1.1)

$$Beta_i = \frac{\max\left(0, y_{methy,\ i}\right)}{\max\left(0, y_{methy,\ i}\right) + \max\left(0, y_{unmethy,\ i}\right) + \alpha}.$$

, where $y_{methy,\ i}$ is the methylated probe intensity, and $y_{unmethy,\ i}$ the unmethylated probe, for i-th CpG site. α is the constant, usually set to 100 (Du et al., 2010). By definition, beta values are always between 0 and 1.

1.2. Epigenetic clocks and their uses

It was only recently since researchers started intensely studying the strong relationship between human aging and epigenetics (Alisch et al., 2012; Bell et al., 2011; Bocklandt et al., 2011; Boks et al., 2009; Bollati et al., 2009; Christensen et al., 2009; Rakyan et al., 2010). The first demonstration of an age predictor was built applying a combination of EWAS and Lasso (least absolute shrinkage and selection operator) penalized regressions to DNA methylation data from saliva samples (Bocklandt et al., 2011). It was also around the same time when robust high-dimensional penalized regression models were applied to human epigenetic methylation data in search of accurate epigenetic aging clocks. I will discuss the basics of Lasso and the elastic net penalized regressions in the following sections. The first of elastic net framework was applied to human blood tissues and was subsequently recognized as the "Hannum clock," which consists of 71 selected CpG sites (Hannum et al., 2013). This clock achieved, in test set of a separate cohort, a correlation of 91% between age and predicted age, and an error of 4.9 years. It was later demonstrated that aging

clocks could be built in almost all human tissues. Horvath lab validified the fact in large multi-tissue human DNA methylation data sets, and coined what would later become one of the most widely recognized human epigenetic aging clocks (S. Horvath, 2013). This multi-tissue epigenetic aging clock, selecting more markers than the Hannum clock, is a linear combination of 353 CpG sites. Contrary to what the name suggests, the aging clocks have been used to infer individuals' rates of aging and the effect of diseases on such rates, rather than simply predicting one's age. The original research work in this dissertation primarily focuses on a unique data set collected from multiple mammalian species. In following chapters, we present prediction and inference models built for mammal species life history traits, including species maximum lifespan, gestation time, and time to sexual maturity, with emphasis on maximum lifespan.

1.3. Dissertation structure

In this introduction chapter, I have briefly described the background of epigenetic aging clocks, which utilize regularized regressions to predict outcome variable, age. In the following chapters of this dissertation, I first describe theoretical groundwork of existing regularized regression models in Chapter 2. Regularized regressions are essential to building mathematical models in high-dimensional data in which the number of columns greatly exceeds the number of rows. These introduced regularized regressions are deeply embedded in the applied as well as original methodology of Chapters 3, 4, and 5 in this dissertation. In Chapter 3, we describe various mammalian history traits prediction models that are trained on the DNA methylation data. In Chapter 4, we present marginal inference statistical models for evaluating individual CpGs sites in relations to the mammalian life history traits. Our analyses place great emphasis on one of such traits, species maximum lifespan, defined as the maximum innate potential of lifespan given any

animal species. Although naturally correlated, maximum lifespan differs from life expectancy, which measures the average lifespan in a population, taking into account diseases and accidental deaths. In Chapter 5, we introduce a novel methodology in the form of a Poisson graphical model for RNA-seq data. Various genetic data types, such as transcriptome (RNA data), epigenome (DNA methylation), proteome data, are crucial for future multiomics data integration, in order to have comprehensive understandings of the species life history traits. In Chapter 6, we address both the current and new statistical methods and outline possible future work that will compare and potentially improve these algorithms.

CHAPTER 2

2. Methodology of regularized regressions

As described in earlier sections, DNA methylation data have tens or hundreds of thousands of variables. Thus, fitting multiple linear regressions using all CpG sites would be overfitting, yielding poor performance in test data. Furthermore, most data sets have much more variables than samples (high-dimensional). This property dictates that the standard linear regression framework would have rank deficiency, producing non-singular solutions. Therefore, in cases of continuous numerical outcome variables, such as chronological age, researchers have been focusing on regularized regression frameworks. Regularized regressions serve as appropriate prediction models, and, for some models, adequate variable selection models. In this section, I will discuss the Lasso, ridge, broken adaptive ridge (BAR), and elastic net regressions, and some of their applications that I have implemented in the field of genetics.

2.1. Lasso and ridge regression

Lasso and Ridge regressions are some of the earliest and most well-known regularized regressions. Lasso was proposed by Tibshirani (Tibshirani, 1996). Lasso imposes an ℓ_1-penalty on a standard multiple linear regression, originally written as,

(Equation 2.1)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{N} \frac{1}{N}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t,$$

where y_i is the i-th observation of sample size of N, $\boldsymbol{x}_i^T$ is the vectorized i-th sample of length p, $\boldsymbol{\beta}$ is the coefficient vector of all p features (variables), and t is a tuning parameter for the shrinkage.

Here the second constraint term of (Equation 2.1, $\|\boldsymbol{\beta}\|_1 \leq t$, ensures that the ℓ_1-norm of all coefficients are shrunk to no larger than t. In practice, (Equation 2.1 is solved in its well-known Lagrangian form,

(Equation 2.2)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{N} \frac{1}{N}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1 .$$

Another popular penalization technique, ridge regression (Hoerl & Kennard, 1970), minimizes the residual sum of squares subject to a bound on the ℓ_2-norm of coefficients. The ridge estimator is directly written as,

(Equation 2.3)

$$\widehat{\boldsymbol{\beta}}^* = [\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}; k \geq 0,$$

where $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{Y}$ is the outcome variable vector, and k is the tuning parameter for the ridge estimator. Note that this matrix form solution is equivalent to being written as the solution for ℓ_2-penalized least square loss function,

(Equation 2.4)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{N} \frac{1}{N}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_2 .$$

The significance of (Equation 2.4 is that it provides a closed-form matrix solution. As $\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}$ is always symmetrical, and especially in low dimensional data sets much smaller than $\boldsymbol{X}$, it can be solved more efficiently via Cholesky decomposition, while Lasso is often estimated by iteratively solving Karaush-Kuhn-Tucker (KKT) condition. However, in contrast to Lasso, ridge regression does not shrink variables to zero, undermining its variable-selection purposes. Consequently, Lasso became widely used in many studies for which ridge penalization is not suited.

Other regularizations have been developed, aiming to achieve an estimation with the oracle property (Fan & Li, 2001). Briefly, oracle property is defined as an estimator that asymptotically

converges to the maximum likelihood estimator (MLE) based on only the true support. In the sense of the variable selection problems, the true support would be considered the unknow set of variables that are truly associated with the outcome. To this end, a few additional ℓ_1-penalization methods have been developed, such as adaptive lasso (Zou, 2006), SCAD (Fan & Li, 2001), and MCP (Zhang, 2010). A more recent regularization method, broken adaptive ridge, aims to improve on these methods' grouping properties for correlated variables (Dai, Chen, Sun, Liu, & Li, 2018; H. Zhao, Sun, Li, & Sun, 2018). I have published a graphical network method paper that utilities its variable selection and superior grouping properties for a discrete RNA-seq data (C. Z. Li, Kawaguchi, & Li, 2021). This graphical model offers a Poisson distribution assumption solution to constructing graphical networks for discrete data sets, such as RNA-seq data. A necessary step for this sparse graphical modeling is regularization, and BAR penalization fits adequately in this algorithm. Nevertheless, elastic net remains the most popular framework for epigenetic aging clocks and a few other DNA methylation predictors, largely due to its speed and robust performance in p>>n ultra-high dimensions.

2.2. Elastic net regularization

In light of searching for a simultaneous prediction and variable selection model that provides towards-zero shrinkage, based on earlier sections, one would consider either Lasso or BAR. However, in practice, Lasso has some pitfalls, including its limitations in p>n cases, variable grouping properties, and inferiority to ridge regression in n>p cases (Zou & Hastie, 2005). Poor grouping in Lasso manifests as its tendency to select only one variable from a group in which variables are highly correlated. This effect could cause Lasso to accidentally drop a true signal variable in favor of another correlated variable. While BAR enjoys the oracle property, it is

computationally expensive for data sets with over thousands of features, due to its iterative ridge regression estimations. The elastic net model is a hybrid penalized model between Lasso and Ridge regressions, and its Gaussian family form is specified as follows,

(Equation 2.5)

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\lambda\left[\frac{(1-\alpha)\|\boldsymbol{\beta}\|_2^2}{2}+\alpha\|\boldsymbol{\beta}\|_1\right],$$

where $\boldsymbol{\beta}$ is the vector of non-intercept variable coefficients, $\boldsymbol{x}_i \in \mathbb{R}^p$ is a vector of i-th observation, and p is the number of CpG probes used in this framework. λ is the Lasso tuning parameter, and α is the ridge penalization parameter. One of the biggest advantages of elastic net is that it performs better than Lasso in p>>n cases (Zou & Hastie, 2005). Furthermore, it handles variable grouping effects more desirably. When several variables are highly correlated each other, Lasso tends to select one of them and ignores the rest, while elastic net might include several most relevant variables from the group. This property is important to our research, as we do not want to leave out important markers that are potentially associated with outcome variable.

2.3. Broken adaptive ridge (BAR) regularization

In addition to Lasso and Elastic net, we describe an alternative multivariate regularization model in this chapter, the broken adaptive ridge (BAR) estimator (Dai et al., 2018). Due to the fact the framework has only been proposed and tested in the recent year, we are yet to have a highly optimized programming package to implement the algorithm at run-times comparable to those of Elastic net (R package: glmnet) and Lasso. Nevertheless, the methodological development of BAR has been advanced to a greater extent of areas in genetics, such as survival Cox model (Kawaguchi,

Suchard, Liu, & Li, 2020), competing risks models (Kawaguchi, Shen, Suchard, & Li, 2021), and regularized graphical models (C. Z. Li et al., 2021) to be discussed in Chapter 5.

BAR estimator takes on an iterative process, which needs to start with initial estimated values. These initial estimates are from a standard ridge regression (Hoerl & Kennard, 1970), defined here as,

(Equation 2.6)

$$\widehat{\boldsymbol{\beta}}^{(0)}(ridge) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} -2l_n(\boldsymbol{\beta}) + \lambda_{ridge}\|\boldsymbol{\beta}\|_2,$$

where $-2l_n(\boldsymbol{\beta})$ can be any objective function, such as a log likelihood function for a generalized regression, or least squared formula for a linear regression, $\|\boldsymbol{y} - \boldsymbol{\beta}\|^2$. For all formula in this section, $\boldsymbol{\beta}$ is the coefficient vector of length p. Superscript (k) denotes the k-th iteration's coefficient estimate.

BAR estimator offers flexibility in regularization weighting for different objectives, such as fused broken adaptive ridge estimator and broken adaptive ridge trend filter (Dai et al., 2018). For simplicity, we focus on variable selection, setting the coefficient weighting vector to $\boldsymbol{d}_j = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the standard basis vector with jth component equal to one. This yields a variant of BAR algorithm that iteratively searches for estimates which regularize by its L_2 norm weighted by its L_2 norm from the last iteration elementwise, formally,

(Equation 2.7)

$$\widehat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_{BAR} \sum_{j=1}^{p} \frac{\beta_j^2}{(\hat{\beta}_j^{(k-1)})^2} \right\}.$$

In this case the final BAR estimator is then defined as,

(Equation 2.8)

$$\widehat{\boldsymbol{\beta}}(BAR) = \lim_{k\to\infty} \widehat{\boldsymbol{\beta}}^{(k)}.$$

Heuristically, to complement ridge regression's non-zero shrinkage, the weighting term in BAR's regularization term's denominator, $\hat{\beta}_j^{(k-1)}$ (Equation 2.8), forces the small coefficient estimates to become smaller. The numerator acting has to be much smaller than its fixed denominator to satisfy the regularization penalty. Asymptotically, BAR shrinks small ridge terms to zero (Equation 2.8). Chapter 5 describes an application of BAR estimator in the context of Poisson graphical model for discrete high-dimensional data.

CHAPTER 3

3. Joint modeling for DNA methylation data

This chapter describes a few research objectives achieved in the context of regularized regression and classification model applications in DNA methylation data. The general aims are to identify epigenetic markers directly involved in lifespan and aging, and to attempt to describe underlying mechanisms of these involvements. While the next chapter focuses more on marginal statistical model application, this chapter's joint modeling research relies on multivariate statistical frameworks that take into account all variables simultaneously. This includes multivariate classifiers such as tissue random forest predictors and multivariate regressions like elastic net. Marginal modeling strategies seek to assess marginal effects in individual CpGs sites. Joint models do not necessarily trump the benefits of marginal frameworks, depending on the context. For example, marginal modeling of markers is valuable to anit-aging intervention studies and possibly future human clinical trial studies. They inform researchers with p-values and confidence intervals that are necessary for assessing the strength of statistical effects.

We leveraged a novel DNA methylation data collection from not just human, but more than 348 mammalian species, enabled by a novel mammalian DNA methylation array (Arneson et al., 2021). Consequently, my research interests have expanded beyond human aging clocks. With such a diverse collection of mammalian species, one is able to directly study the secret to long lifespan. Some of the most important response variables in multi-species mammalian epigenetic aging studies include species maximum lifespan, organism chronological age, sample tissue type, and sample sex annotation. Most importantly, maximum lifespan of a species, in particular, is the oldest that individuals can survive, reflecting the genetic limit of longevity in an ideal environment. The maximum lifespan of humans and other mammals appears to be fixed and subject to natural

constraints (Dong, Milholland, & Vijg, 2016). We recognize that the molecular mechanisms underlying these constraints remain poorly understood (Austad, 2010; de Magalhaes, Costa, & Church, 2007), despite prior studies correlating maximum lifespan with specific molecular processes and life history strategies (Gorbunova & Seluanov, 2009; Harper, Salmon, Leiser, Galecki, & Miller, 2007; Tian et al., 2019). Some researchers have suggested that epigenetic mechanisms may play a role in controlling lifespan and aging (Booth & Brunet, 2016; de Magalhaes, 2012; Lowe et al., 2018; Mayne, Berry, Davies, Farley, & Jarman, 2019; Mitteldorf, 2016; Rando & Chang, 2012; Sen, Shah, Nativio, & Berger, 2016; Wilkinson et al., 2020; J.-H. Yang et al., 2019). The role of epigenetics in mammalian aging is underscored by recent studies demonstrating age reversal through (transient) epigenetic reprogramming with Yamanaka factors (Gill et al., 2021; Y. Lu et al., 2020; Ocampo et al., 2016; Rodríguez-Matellán, Alcazar, Hernández, Serrano, & Ávila, 2020; Sarkar et al., 2020; Takahashi & Yamanaka, 2006).

While the data set continues to expand, I will list a few underpinnings discovered in maximum mammalian lifespan and other life history traits using DNA methylation profiles from 348 mammalian species, from 25 taxonomic orders including primates, rodents, bats, cetaceans, and marsupials. The life history traits data are from a current of anAge database (de Magalhaes et al., 2007). I successfully developed methylation-based predictors of time-related life history traits: maximum lifespan, gestation time, and age at sexual maturity across therian mammalian species.

3.1. Methodology

3.1.1. Data collection

All data were generated using the mammalian methylation array (HorvathMammalMethylChip40) (Arneson et al., 2022) which provides high sequencing depth of highly conserved CpGs in

mammals. Out of 37,492 probes (cytosines) on the array, 35,989 probes were chosen based on high levels of sequence conservation within mammalian species (Arneson et al., 2022). The particular subset of species for which each probe is expected to work is provided in the chip manifest file which can be found at the NCBI Gene Expression Omnibus (GEO) as platform GPL28271, and on our Github webpage. The SeSaMe normalization method was used to define beta values for each probe and to calculate detection p values (Zhou, Triche, Laird, & Shen, 2018). DNA methylation data generated by this array can be used to accurately classify sample species, sex, and tissue in randomly selected test sets (or random forest out-of-bag sets) (C. Li et al., 2021). We analyzed methylation data from 348 mammalian species representing 25 out of 26 taxonomic orders. (C. Li et al., 2021). The only order not represented was the marsupial order Peramelemorphia. DNA was derived from 59 different tissues and organs including blood, skin, liver, muscle, and brain regions (C. Li et al., 2021). Supplementary Information contains details on all the data sets that we have used to conduct analyses.

3.1.2. Life history traits and anAge database

The high accuracy of the epigenetic estimator of maximum lifespan is a testament to the success of a decade-long effort of biologists and the anAge database (de Magalhaes et al.) to establish this elusive phenotype. For several species, maximum lifespan was not available in anAge. In this case, we updated the results based on literature searches. For select species, we used a K=1 nearest neighbor predictor to impute values. For this reason, our KNN based predictor of life history traits is biased. To enhance the reproducibility of our findings we include our updated version of the anAge database (de Magalhaes et al.) (C. Li et al., 2021).

3.1.3. Multivariate estimators of maximum lifespan

For most species, relatively few animals informed the determination of maximum lifespan, which may bias this life history trait (Ronget & Gaillard; Vaupel). To account for the fact that the maximum lifespan of humans and mice was established on the basis of many studies while the maximum lifespan of other mammalian species was based on fewer animals, we corrected the maximum lifespan value of the remaining species by multiplying it by 1.3. This adjustment step assumes that each maximum lifespan estimate reported in anAge underestimates the true value by 30 percent in all species except for humans and mice. We applied the same adjustment step in our universal mammalian clock project (Ake T. Lu et al., 2021). In addition, in the final model fitted to all species as a training set, we calibrated the predictor by the mean and standard deviation, similar to those of biomarker, to match those of the observed lifespan (Ake T Lu et al., 2019). This correction was only used for our multivariate estimator of maximum lifespan, e.g. we did not use it in our EWAS.

We used elastic net regression to build different multivariate predictors of maximum lifespan, gestation time, and age at sexual maturity (Zou & Hastie, 2005). We chose a more data-driven and less human-mice-centered method for variable selection, i.e. CpG screening. To build a model on the basis of CpGs that are present/detectable in most species, we restricted the analysis to CpGs with significant median detection p-values (false discovery rate<0.05) (Y. Benjamini & Hochberg, 1995) in 85% of the species. This resulted in a lower-dimensional dataset consisting of 17,032 CpGs.

We employed two strategies for building lifespan predictors. The first strategy ignored tissue type. Here, all tissue samples from a given species were averaged resulting in a single observation per species. The second strategy formed average values for each stratum defined by

tissue type and species. For example, this analysis formed an average value for human blood (considered as one stratum). The second approach allowed us to study the influence of tissue type on lifespan predictions. This second strategy shows similar prediction correlations in all three life history traits. To arrive at unbiased estimates of the predictive accuracy of lifespan and other predictors, we used a leave-one-species-out (LOSO) cross-validation analysis that iteratively trained the predictive model on all but one species. Next, the predictor was applied to the observations from the left-out species. By cycling through the species, we arrived at LOSO estimates for each species. As a sensitivity analysis, we also conducted a leave-one-clade-out analysis as described below.

3.1.4. Leave one clade out cross validation

In order to evaluate the taxonomic robustness of the DNAm predictions (section entitled **DNAm-based predictors out-perform phylogeny-based predictors**), we iteratively left out taxonomic orders as test sets, in addition to conducting leave-one-species out (LOSO) analysis. Setting aside entire orders as test sets serves to validate the predictor's performance when given taxonomically (phylogenetically) different species. We could not carry out a leave-one-taxonomic order out cross validation analysis because of the highly skewed distribution of animals across taxonomic order: Rodentia contained 27% of all species while many other orders contained fewer than 3% of the species. To address this challenge, we modified the leave-one-order-out analysis by leaving out all but two species as a test set (corresponding to the minimum and maximum lifespan) in a few taxonomic orders with more than 20 species (Rodentia, Artiodactyla, Chiroptera, Primates, Carnivora, and Eulipotyphla). The two species per large order kept in the training set act as a "counter-weight", challenging the predictor to guess the lifespan for the rest of the taxonomic order

given this limited information. All species in a small order are left out in its entirety as test sets. For example, taxonomic orders Dasyuromorphia, Microbiotheria, Microbiotheria, Sirenia, and Tubulidentata were represented by a single species. We refer to the resulting cross validation scheme as leave-one-clade-out (LOCO) analysis. A predictor too reliant on neighboring species with similar lifespan in proximity, such as the tree-based KNN, would under-perform in such training-test separation scheme. In addition, due to the fact that we imputed a number of species' missing lifespan observations using KNN, lifespan estimates naturally favor k-NN. Thus, in this analysis only, we use the original anAge database (de Magalhaes et al.) that did not contain any imputed values. It is evident that KNN lifespan predictor, despite having acceptable prediction correlation, gives constant and off-center predictions for entire taxonomic orders (C. Li et al., 2021). For any test set, It tends to find the "nearest" species to be the two species given in LOCO (or some species in a neighboring order for small test-set orders), resulting in the same estimate for every member of that taxonomic order. Thus, such an algorithm is undesirable when applied to dissimilar species or clades.

3.2. Results

3.2.1. Multivariate predictors of life history traits

We fitted three separate penalized regression models to predict log-transformed values of maximum lifespan, gestation time, and age at sexual maturity for each species. We obtained the species values for these traits from the current version of the anAge database (de Magalhaes et al.; C. Li et al., 2021). The resulting epigenetic predictors exhibited a high level of accuracy according to leave-one-species-out (LOSO) cross-validation, e.g., the predicted log maximum lifespans were highly correlated with those documented in anAge (Pearson's correlation R = 0.89, Figure 3.1a &

3.1b). Actual log gestation time, which is easier to measure accurately than maximum lifespan, exhibited an even higher correlation with predicted log-gestation time (R = 0.96, Figure 3.1c). Interestingly, the epigenetic estimator of (log-transformed) age at sexual maturity exhibited a relatively lower correlation of R = 0.85 with documented measurements (Figure 3.1d). This may partly reflect that age at sexual maturity is far more malleable than gestation time, depending on food availability and various ecological/environmental factors. An alternative 70%-30% training-test random separation scheme yields similarly high correlations for log maximum lifespan in both the training and test sets (training set, R = 0.98, Figure 3.2a; test set, R = 8.8, Figure 3.2b).

We hereafter refer to the predicted maximum lifespan, in units of log years, as epigenetic maximum lifespan or DNA methylation (DNAm) maximum lifespan. The same nomenclature applies to other DNAm-based estimates of life history traits. We carried out two analyses to study the relationship between epigenetic maximum lifespan and chronological age of the individuals of species sampled. First, we built a separate maximum lifespan predictor using only samples obtained from animals that were younger than their species' average age of sexual maturity and younger than 5, and this had acceptable correlation in lifespan prediction (R = 0.68, Figure 3.3), even though the restriction of age resulted in fewer species (n = 122) being available for this analysis.

Second, we applied the final lifespan predictor model to individual animal samples. Once a final model had been fitted to all species-wise averaged data, the regression model coefficients were frozen. Despite the fact that the predictor was intended for predicting species level lifespan on a log scale, we applied coefficients in an attempt to predict individual samples' lifespan. We show that although predicted maximum lifespans for individual samples can vary and correlate with chronological age in a few species (e.g. naked mole rat skin, brown rat blood, sheep, human

blood) (Figure 3.4), the vast majority of the individual sample predictions remain incredibly stable in most species-tissue strata (Figure 3.4). Epigenetic maximum lifespan also depends on tissue type (Figure 3.5 and Figure 3.6). In humans, the final epigenetic maximum lifespan estimates are 97.7 years for blood, 94.5 for epidermis, 77.9 for skin, and 49.8 for cerebral cortex (C. Li et al., 2021).

Considering the high correlation of maximum lifespan and adult weight, we examined the potential confounding effects of average adult weight on the performance of our model. The LOSO estimate of log transformed maximum lifespan was moderately correlated ($R = 0.54$, $P < 2.2 \times 10^{-16}$) with the weight adjusted maximum lifespan. A multivariate regression model (dependent variable log of maximum lifespan) revealed that log adult weight (Wald test $P = 1.3 \times 10^{-6}$) is a less significant covariate than the log transformed epigenetic maximum lifespan ($P < 2 \times 10^{-16}$).

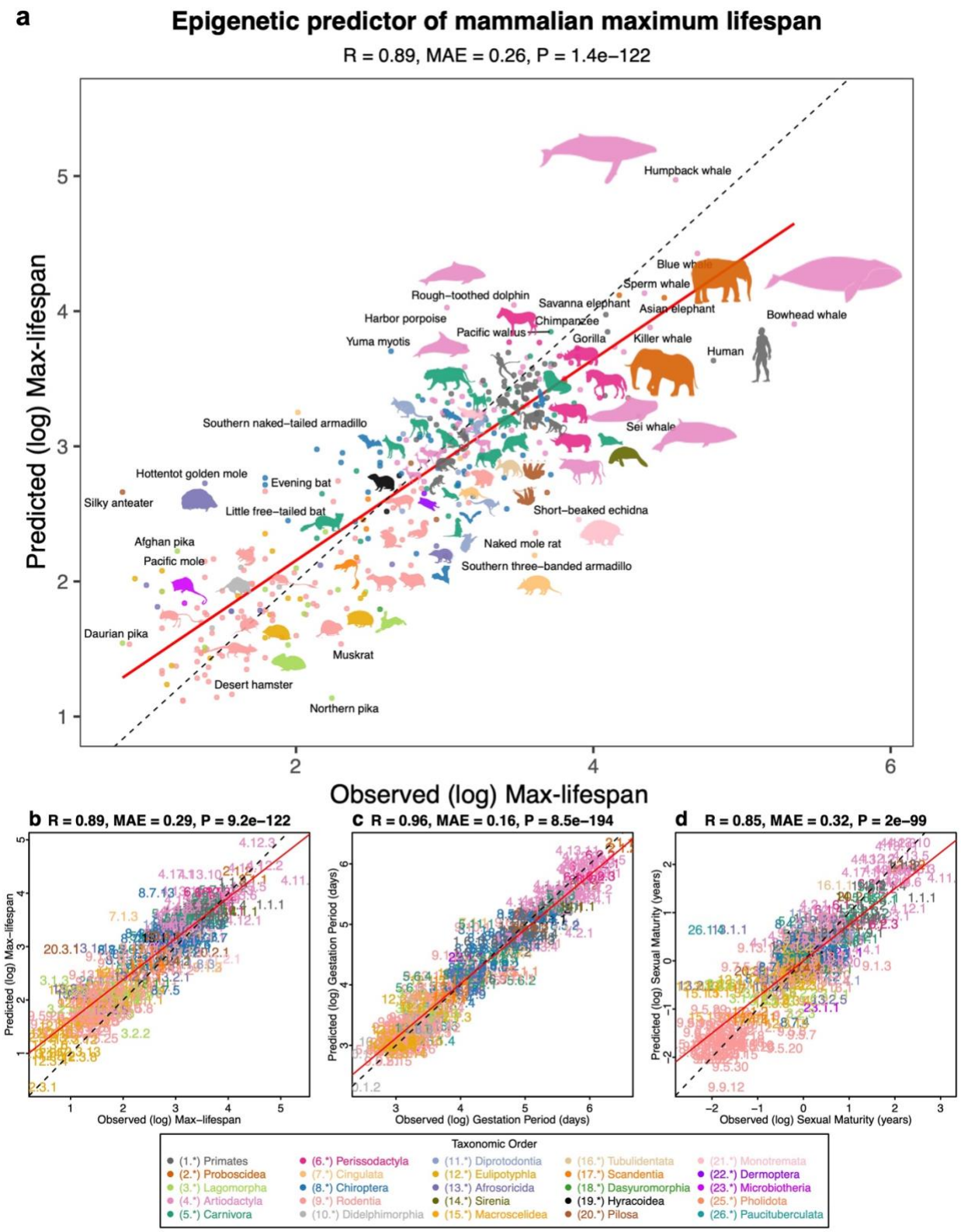
a
Epigenetic predictor of mammalian maximum lifespan
R = 0.89, MAE = 0.26, P = 1.4e−122
Predicted (log) Max-lifespan
Observed (log) Max-lifespan
Humpback whale
Blue whale
Sperm whale
Rough−toothed dolphin
Savanna elephant
Asian elephant
Bowhead whale
Harbor porpoise
Chimpanzee
Pacific walrus
Gorilla
Killer whale
Yuma myotis
Human
Southern naked−tailed armadillo
Sei whale
Hottentot golden mole
Evening bat
Silky anteater
Little free−tailed bat
Short−beaked echidna
Afghan pika
Naked mole rat
Pacific mole
Southern three−banded armadillo
Daurian pika
Muskrat
Desert hamster
Northern pika
b R = 0.89, MAE = 0.29, P = 9.2e−122
Predicted (log) Max−lifespan
Observed (log) Max−lifespan
c R = 0.96, MAE = 0.16, P = 8.5e−194
Predicted (log) Gestation Period (days)
Observed (log) Gestation Period (days)
d R = 0.85, MAE = 0.32, P = 2e−99
Predicted (log) Sexual Maturity (years)
Observed (log) Sexual Maturity (years)
Taxonomic Order
(1.*) Primates
(2.*) Proboscidea
(3.*) Lagomorpha
(4.*) Artiodactyla
(5.*) Carnivora
(6.*) Perissodactyla
(7.*) Cingulata
(8.*) Chiroptera
(9.*) Rodentia
(10.*) Didelphimorphia
(11.*) Diprotodontia
(12.*) Eulipotyphla
(13.*) Afrosoricida
(14.*) Sirenia
(15.*) Macroscelidea
(16.*) Tubulidentata
(17.*) Scandentia
(18.*) Dasyuromorphia
(19.*) Hyracoidea
(20.*) Pilosa
(21.*) Monotremata
(22.*) Dermoptera
(23.*) Microbiotheria
(25.*) Pholidota
(26.*) Paucituberculata

Figure 3.1 Scatter plots of Leave-one-species-out (LOSO) cross-validation analysis of epigenetic test set predictions. Y-axes show log (base e) transformed estimates of a,b maximum lifespan (in years), c, gestation time (in days), and d, age at sexual maturity (in years). Each species is represented by a number whose integer part denotes the taxonomic order. Each data point number corresponds to a different species and is color-coded according to order (The silhouettes images of animals were acquired from Phylopic database and are under Public domain or Creative Commons license). Numeric values can be found in shorvath/MammalianMethylationConsortium and C. Li et al. (2021). The titles of the panels report Pearson correlation coefficients, median absolute errors (MAE), and p-values. Colors represent taxonomic order annotation consistent with those of other figures. Species appear as designated numbers in scatter plot panels; the corresponding taxonomic orders are annotated in figure legends; the first whole number (number before the decimal separator) part of each mammalian number is assigned in accordance to the corresponding taxonomic order. Red solid line represents the perfect prediction line, and the dotted line represents the fitted linear regression line.

3.2.2. DNAm-based predictors outperform phylogeny-based predictors

Since DNA methylation levels are under strong genetic control, our DNAm-based lifespan predictor may only be capturing DNA sequence differences driven by phylogenetic relationships rather than by determinants of lifespan in each individual species. This, however, does not appear to be the case, as we learned from two separate analyses.

First, we train elastic net regression models of maximum lifespan, our dependent variable, on the basis of both CpG methylation data and indicator variables for all taxonomic orders (both used as covariates in a multivariate model). The elastic net model only selected CpG methylation as significant covariates, suggesting that CpG data explained more of the variation in maximum lifespan than taxonomic indicator variables.

Second, we compared the accuracy of the DNAm lifespan predictor to that of k-nearest neighbor (KNN) regression predictors that used a distance measure defined by the branch length of the phylogenetic tree. The simplest version of the KNN predictor results from a choice of K = 1 which entails that the maximum lifespan of a given species will be predicted by that of its closest neighboring species in the taxonomy.

Phylogeny-based k-NN predictor unsurprisingly performs almost as well as DNAm predictor under leave-one-species-out training-test separation scheme, because the vast majority of mammalian species in our dataset share similar maximum lifespan with neighboring species (Figure 3.7a and Figure 3.7b). This is also evident on the taxonomic family level (Figure 3.7c and Figure 3.7d), as species in the same families have similar lifespan (C. Li et al., 2021). To arrive at an unbiased comparison between methylation-based predictors and the KNN predictors, we used a special, more stringent cross-validation scheme, referred to as leave-one-clade-out (LOCO), analysis (Methods), to challenge the predictors to predict taxonomically different species. A predictor solely reliant on phylogeny (taxonomy) would yield inferior predictions under this training-test scheme. Although the KNN predictors were moderately accurate at predicting log-transformed maximum lifespan (LOCO cross-validation estimate of Pearson correlation R = 0.62 for K = 1 NN predictors, Figure 3.2), they are inferior to the methylation-based predictor (R = 0.73 Figure 3.7c). A KNN predictor with K = 2 neighbors led to a correlation of R = 0.62 while a value of K=3 led to R = 0.57. A closer look at the residuals of the breakdown of the test sets reveals k-NN predictions in large taxonomic orders are constant and off-center, resembling poor guesses based on taxonomically distant species (Figure 3.8).

Overall, KNN predictors based on phylogeny are clearly inferior to DNAm-based predictors according to LOCO cross-validation analyses. The fact that DNAm predictors can predict lifespan in taxonomic orders that were omitted from the training set suggests that DNAm captures an aspect of mammalian lifespan that transcends phylogeny.

3.2.3. Sex differences in predicted lifespan

Once a final model had been fitted to all species-wise averaged data, the regression model coefficients were frozen. We then applied coefficients in an attempt to predict individual samples' lifespan, despite the fact that the predictor was trained to predict species level lifespan on a log scale. The predicted maximum lifespan based on female tissues is highly correlated with that based on male tissues (on a log scale, R = 0.99). Most species showed consistent epigenetic estimates of maximum lifespan in female and male samples (C. Li et al., 2021). Stratifying by tissue type, we observed a significant sex difference in epigenetic maximum lifespan (two-sided Student t-test P < 0.01) in only 18 species (Figure 3.2e), in which all tissues showed female-male difference unanimously. Females were predicted to have a longer maximum lifespan than males in 17 of the 18 species, including humans (Figure 3.2e). Across all species, females have a 1.8% longer predicted epigenetic maximum lifespan than males of the same species.

3.2.4. Lifespan predictor does not simply reflect body mass

We observed that maximum lifespan and average adult weight (body mass) are highly-correlated across species (Figure 3.9a), a finding consistent with previous studies (de Magalhaes et al., 2007). We considered the likelihood that the impressive accuracy of epigenetic lifespan predictors may be due to the confounding effect of average adult weight. This, however, is not the case as the epigenetic predictor of maximum lifespan remains highly correlated with the observed values (Pearson correlation R = 0.56, P = 3.3×10^{-10}, Figure 3.9c) in small species (defined as average adult weight <150 grams in our data) even though adult weight is *negatively* correlated with maximum lifespan in those species (R = -0.21, Figure 3.9b). Overall, this demonstrates that epigenetic maximum lifespan captures information beyond adult weight.

3.2.5. Cross-species classifiers of sex and tissue and other categorical outcomes

We have conducted tissue, species, and sex classification using supervised learning methods, random forest (Liaw & Wiener, 2002) and logistic elastic net (Zou & Hastie, 2005). For the tissue and species classification, random forest achieved an out-of-bag accuracy of over 98.2% and 99.9%, respectively (Table **3.1**). With a random 70-30% training-test separation, logistic elastic net regression was able to predict mammalian sample sex with over 98% accuracy in test data (Table **3.1**). However, even robust methods such as the random forest is subject to bias, such as favoring large categorical groups in unbalanced data set. To counter the fact that some species have much more samples than others, the random forest bootstrap step was slightly modified to draw at a cap of 100 samples from each species. In general, regardless of which robust classification model to use, the mammalian array DNA methylation data can be used to effectively classify sample sex, tissue type, and species.

Table 3.1. Variable Classification by DNA Methylation Data

Classification Variable	Predictor Framework	Method Note	Test set / Out-of-bag Accuracy
Tissue	Random Forest	100 trees*	98.22%
Species	Random Forest	100 trees	99.94%
Sex (Female = yes/no)	Elastic Net		98.53%

Note: *100 trees: random forest was calibrated to use this many decision trees for a reasonable run time; random forest unbiased prediction accuracy estimate is calculated as follows; first, summarize by calculating each category's out-of-bag prediction errors, subtracted by unity, across all trees used.

In addition to multi-species sample sex classifier based on the 40K Mammalian Arrays (Arneson et al., 2022). The classifier has a 10-fold training-test set cross-validation accuracy of 98.6%. We translated this classifier to a new Illumina mouse 320K DNA Methylation array, yielding an accuracy of 97.5% (Table **3.2**).

Table 3.2: Mammalian Array Classifier Performance on 320K Methylation Array

Classification Variable	Predictor Framework	Probe Screening	Accuracy 320K	Accuracy 40K
Tissue	Random Forest	Good quality 1-1	94.01%	97.75%
Species	Random Forest	Good quality 1-1	Close to 100%	99.95%
Sex (Female = yes/no)	Elastic Net	Good quality 1-1	97.54%	98.56%

Note: *Good quality 1-1: prior to model fitting, we subset probe (feature) set to the intersection of Illumina 40K and 320K microarray probes, as well as those high quality probes that performed well in mice calibration data set.

3.3. Additional figures

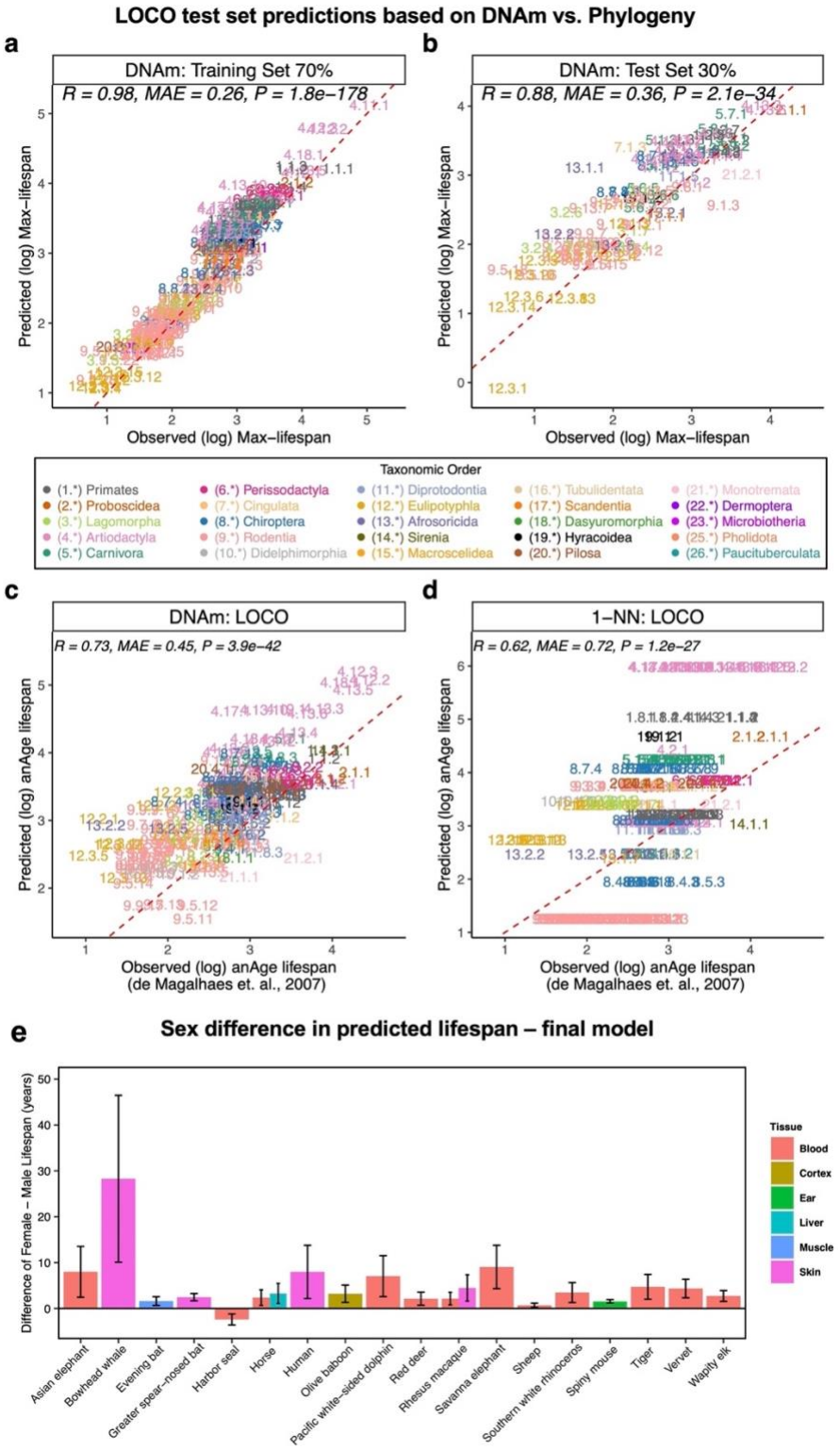

Figure 3.2: DNAm lifespan predictor vs phylogeny-based predictor and sex differences in predicted lifespan. LOCO, leave-one-clade-out, cross-validation analyses of predictors of log (base

e) transformed estimates of maximum lifespan. We compare prediction performance between DNAm elastic net predictors and 1-Nearest-Neighbor predictor (KNN). 1-Nearest-Neighbor predictor utilizes distances from the Mammalian phylogenetic TimeTree (Kumar, Stecher, Suleski, & Hedges, 2017). Panels show **a**, DNAm predictor's test set predictions, **b**, k-NN predictor's test set predictions. In addition, due to the fact that we imputed a number of species' missing lifespan observations with neighboring species, lifespan estimates naturally favor k-NN. Thus, in this analysis only, we use the original anAge database (de Magalhaes et al.), removing species with no maximum lifespan estimates. Panels **b** and **c** report randomly separated training set comprising 70% of species and a test set consisting of the rest 30%, respectively. Panel **e** reports differences between female and male lifespan final model predictions in species in which they show statistical significance. Bars are colored by tissue type as indicated in the legend. For panels **a** and **b**, each data point in the panels corresponds to a different species and is color-coded according to taxonomic order. Red solid line represents the perfect prediction line, and the dotted line represents the fitted linear regression line. Panel **c** reports final DNAm lifespan female vs. male predictions for species in which the predictions differ significantly with a two sample T-test p-value less than 0.01. Error bars represent the 95% confidence interval of two sample mean differences.

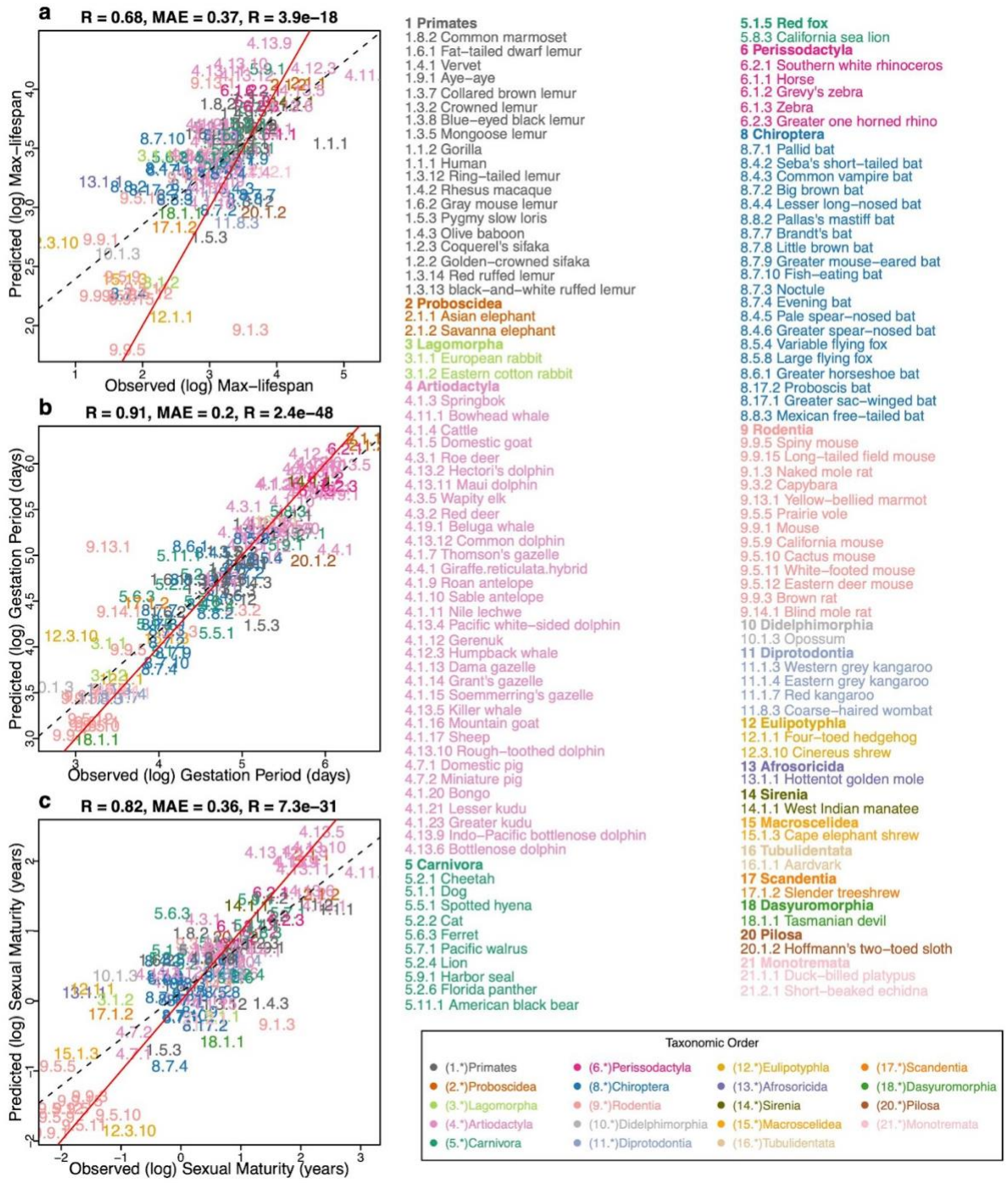

Figure 3.3: Elastic net Predictor Based on Young Samples. Elastic net predictor, Leave-one-species-out analysis, fitted on a subset of all young samples (species n = 119). Young samples are defined as samples whose age is both younger than five years and less than the species' average age at sexual maturation. Feature filtering and Elastic Net tuning parameter set-up is the same as those for Figure 3.1. Three panels show predictors for **a,** log maximum lifespan (in log years), **b,** log-transformed gestation time (in log days), and **c,** log-transformed age at sexual maturity (in log years). As with the Figure 3.1, species appear as designated numbers in scatter plot panels; the corresponding common names and phylogenetic orders are annotated in figure legends; as

indicated by the taxonomic order legend, the whole number (number before the decimal separator) part of each mammalian number is assigned in accordance to the corresponding taxonomic order. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Numeric values can be found in C. Li et al. (2021). Red solid line represents the perfect prediction line, and the dotted line represents the fitted linear regression line.

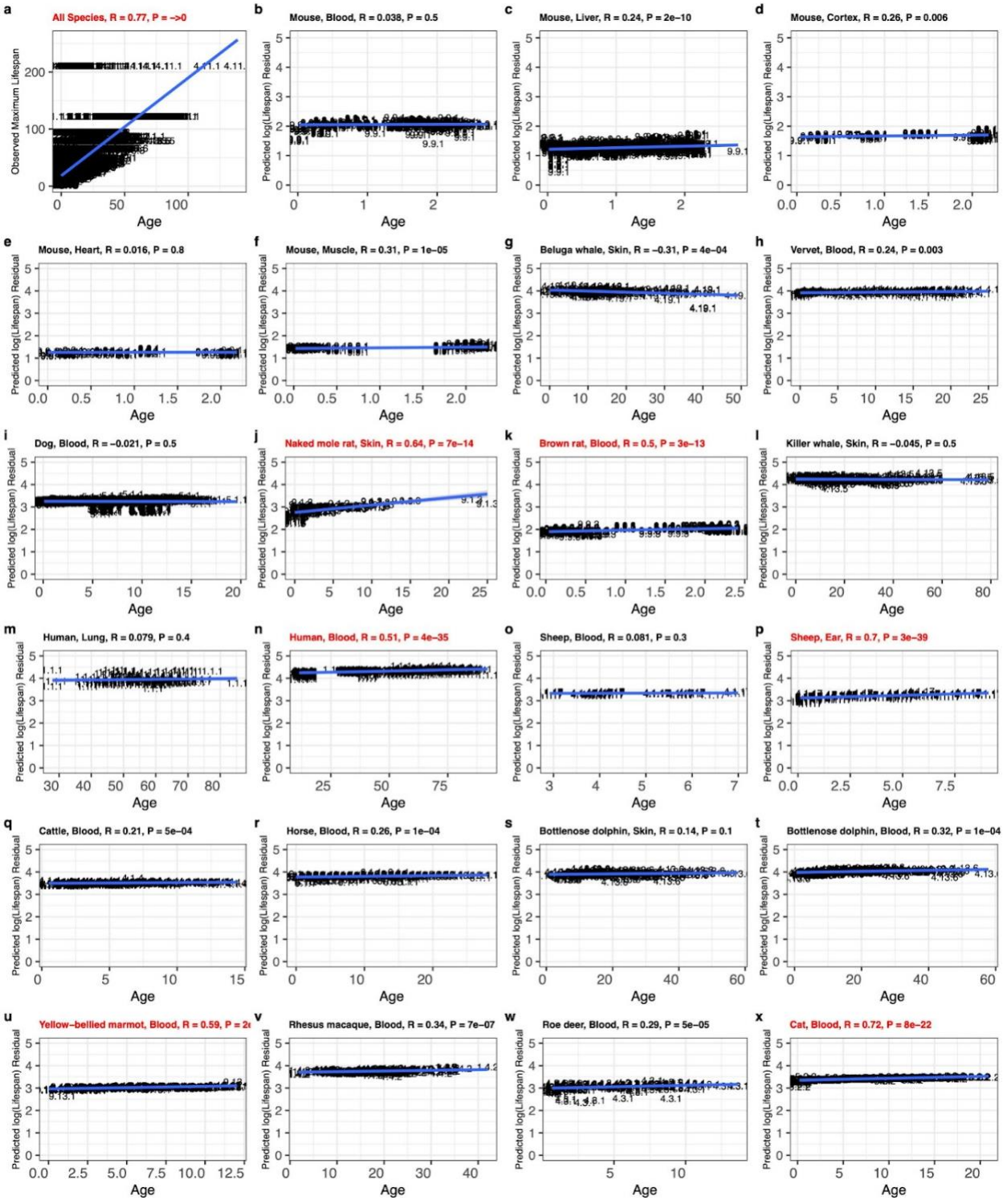

Figure 3.4: Correlation between maximum lifespan predictor and sample chronological ages. Mammalian maximum lifespan predictor, based on averaged species methylation, was used to predict individual sample lifespans. The predicted values are also stratified by species and tissues. Only species with >100 sample sizes are shown. Color scale: pink, female; black, male. To demonstrate natural relations between maximum lifespan and chronological age, panel **a** scatter

plot shows association between maximum lifespan and chronological age of corresponding samples. Each of panels **b–x** show scatter plots of predicted lifespans in log scales vs. chronological age in specific species. Numbers are the mammalian species number consistent with those of other figures. Numeric values can be found in Github repository shorvath/MammalianMethylationConsortium. Shaded areas represent 95% confidence intervals of the simple linear regression line. Colors represent male and female annotation.

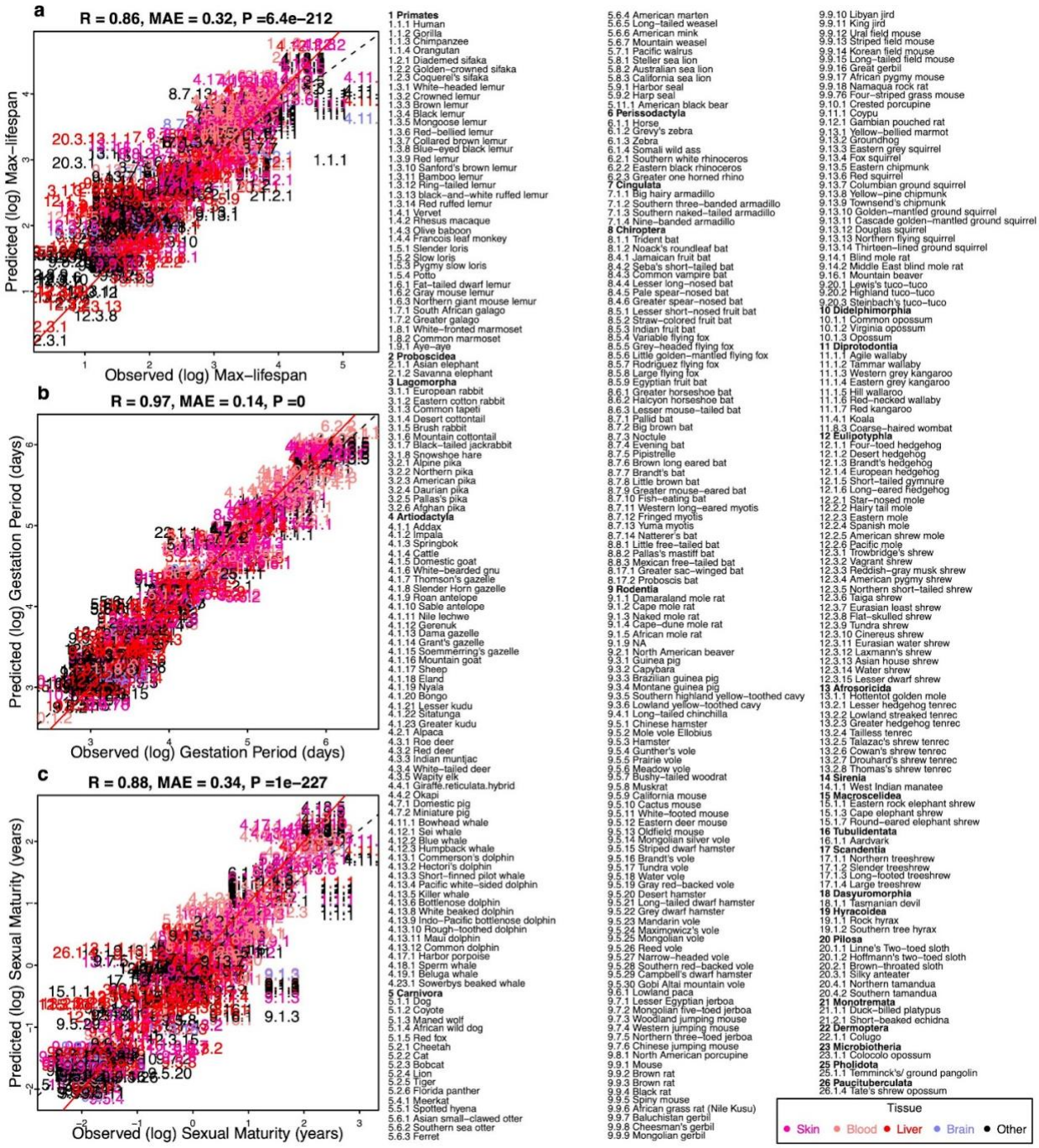

Figure 3.5: Predictors of Species-Tissue Combinations. A penalized joint linear model used to predict species lifespan (Elastic net). Same framework as that of Figure 3.1, except that it distinguishes tissue types. CpG probes are averaged by each species-tissue combination. Different tissues within the same species share the same maximum lifespan, but retain different methylation levels. Three panels show predictors for **a,** log maximum lifespan (in log years), **b,** log-transformed gestation time (in log days), and **c,** log-transformed age at sexual maturity (in log years).

Designated Mammalian numbers in scatter plot panels and the figure legend are the same as those of main Figure 3.1. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Numeric values can be found in Github repository shorvath/MammalianMethylationConsortium. In Figure 3.1, species appear as designated numbers in scatter plot panels; the corresponding common names and taxonomic orders are annotated in figure legends; the whole number (number before the decimal separator) part of each mammalian number is assigned in accordance to the corresponding taxonomic order. Red solid line represents the perfect prediction line, and the dotted line represents the fitted linear regression line.

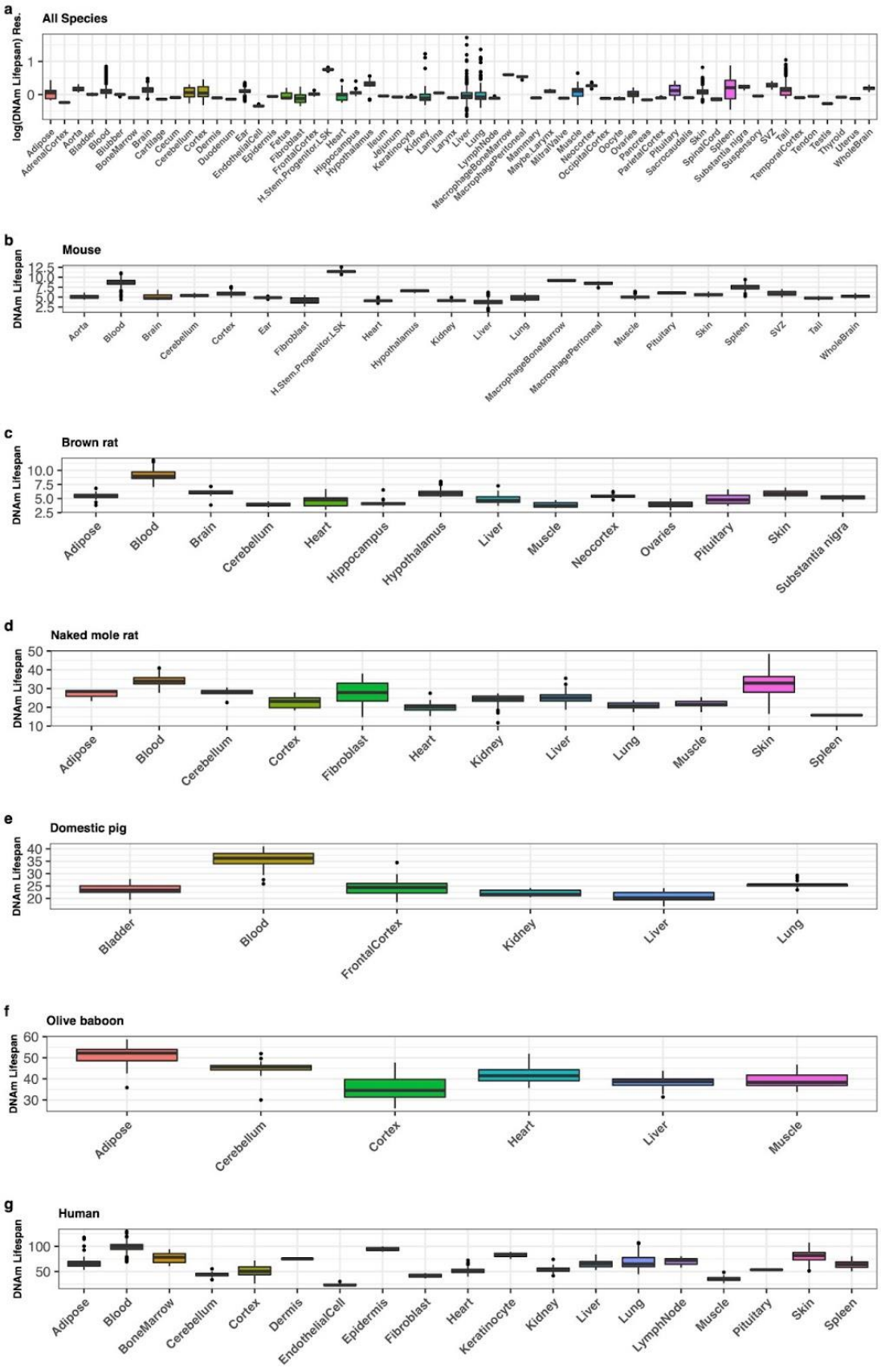
a
All Species
log(DNAm Lifespan) Res.
b
Mouse
DNAm Lifespan
c
Brown rat
DNAm Lifespan
d
Naked mole rat
DNAm Lifespan
e
Domestic pig
DNAm Lifespan
f
Olive baboon
DNAm Lifespan
g
Human
DNAm Lifespan

Figure 3.6: Tissue groups differences in predicted mammalian maximum lifespan. Mammalian maximum lifespan predictor, based on averaged species methylation, was used to predict individual sample lifespans. The predicted values are grouped by sample tissue annotations. Panel **a** shows predicted maximum lifespans (DNAm lifespan) standardized residuals (Res.) by tissue groups in all species and samples; in order to show viewable scales in different species, due to their drastically different lifespans, we evaluated residuals standardized by species (log of predicted maximum lifespan minus log of observed maximum lifespan, results from which are divided by log of observed maximum lifespan of the species to which the samples belong); panel **b–g** show boxplots of predicted lifespans in original scales (DNAm lifespan) by tissue groups; only species with more than 5 tissue types; due to the fact that within-species comparisons require no re-scaling, predicted lifespans (in years) are shown in these panels; Tissue type "H.Stem.Progenitor.LSK" stands for "LSK Progenitor Hematopoietic Stem cells"

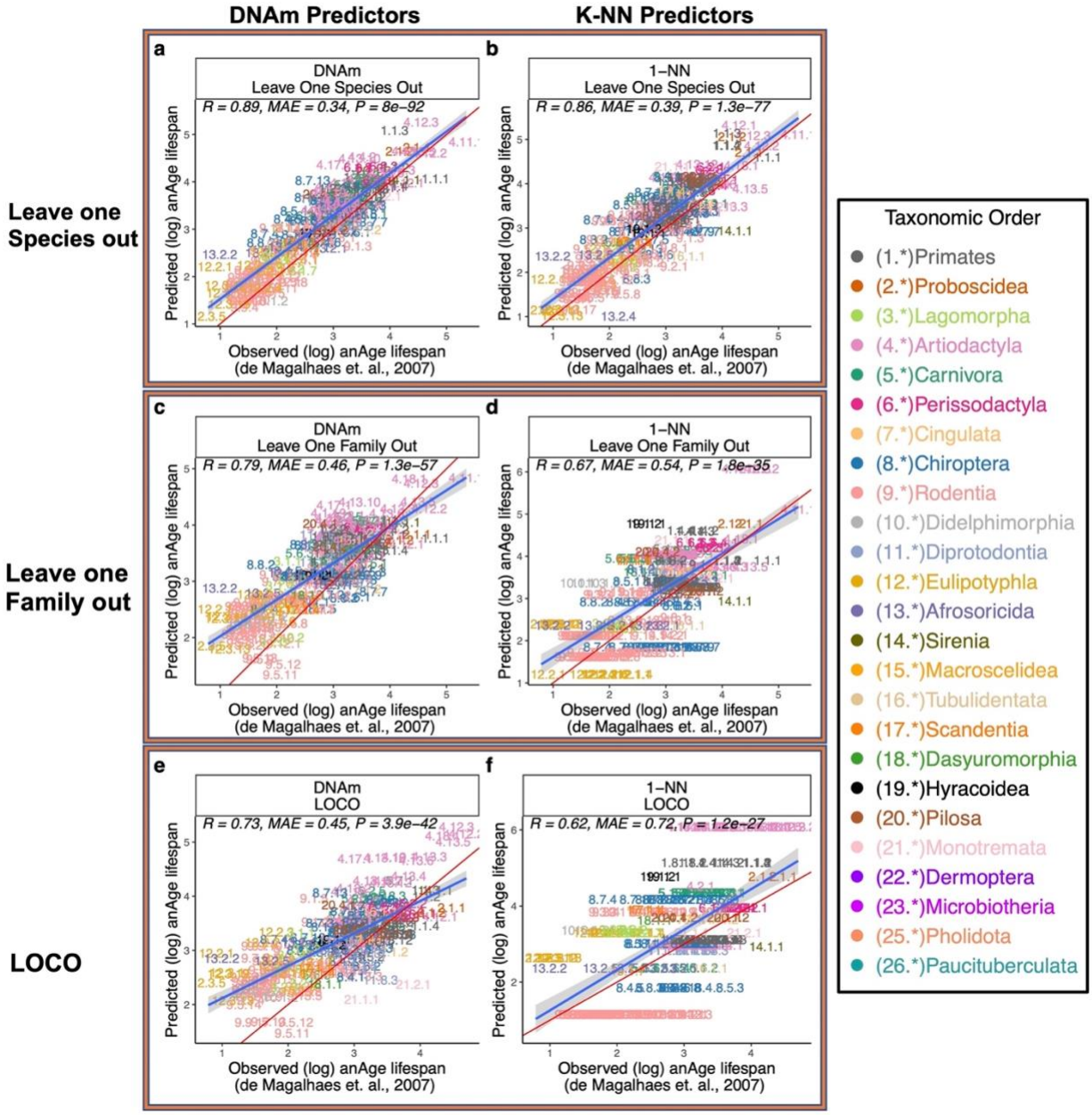

Figure 3.7: Overall Comparisons between DNAm lifespan predictors and Phylogeny-based Predictors. Various training-test validation analyses of predictors of log (base e) transformed estimates of maximum lifespan. We compared prediction performance between DNAm elastic net predictors and 1-Nearest-Neighbor predictor (KNN). 1-Nearest-Neighbor predictor utilizes distances from the Mammalian phylogenetic TimeTree (Kumar et al., 2017). Results under different training-test separation methods are shown in panels **a**, **b**, DNAm and k-NN predictors test set predictions under leave-one-species-out (LOSO) training-test separation scheme; **c**, **d**, DNAm and k-NN predictors test set predictions under leave-one-family-out training-test separation; **e**, **f**, DNAm and k-NN predictors test set predictions under leave-one-order-out

training-test separation; **g**, **h**, DNAm and k-NN predictors test set predictions under leave-one-clade-out (LOCO) training-test separation. LOCO (leave-one-clade-out) is defined as, for orders with more than 20 species (Rodentia, Artiodactyla, Chiroptera, Primates, Carnivora, and Eulipotyphla), leaving out all member species except the longest-living and shortest-living species. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Numeric values can be found in Github repository shorvath/MammalianMethylationConsortium. Shaded areas represent 95% confidence intervals of the simple linear regression line. E).

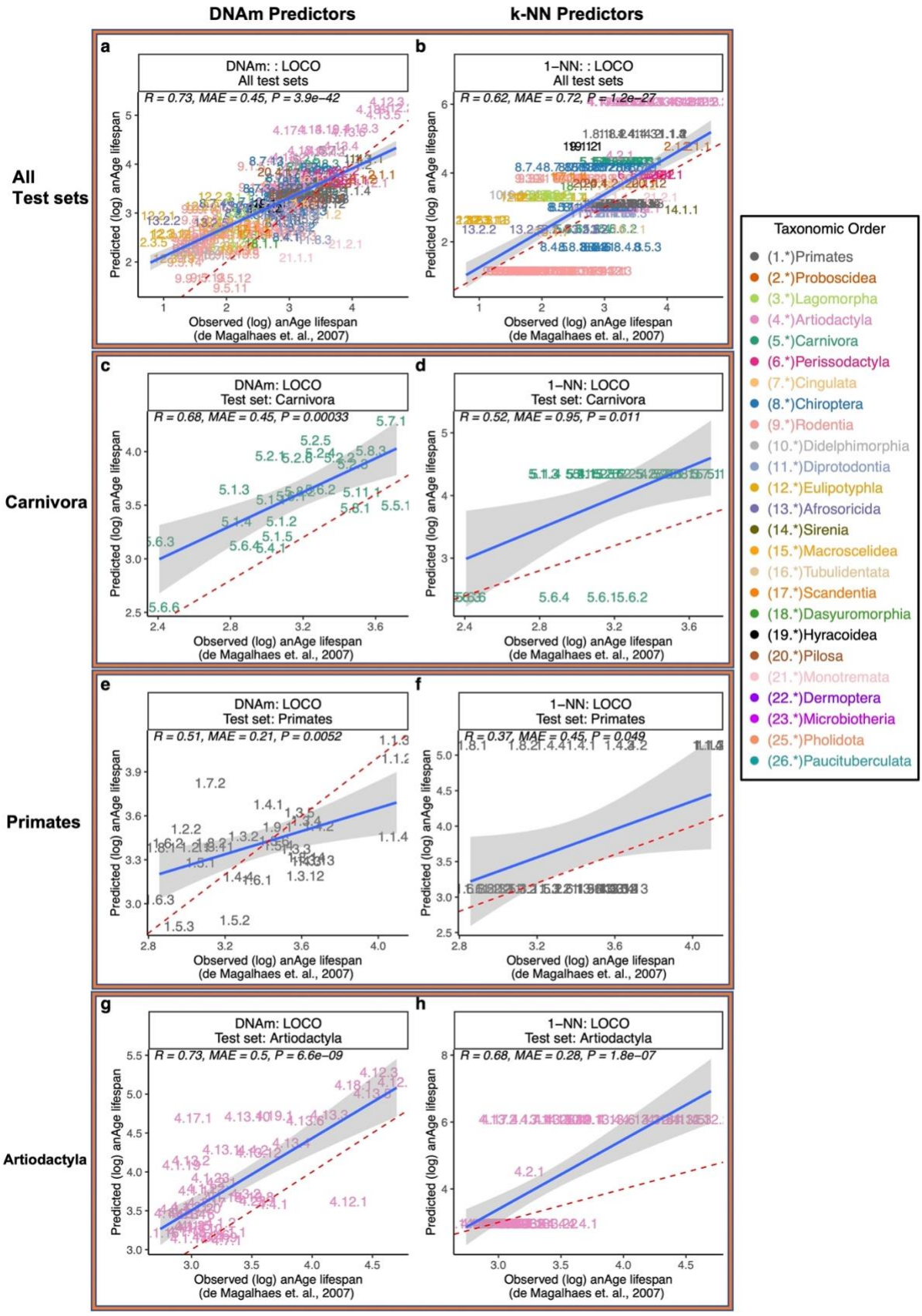
DNAm Predictors
k-NN Predictors
All Test sets
Carnivora
Primates
Artiodactyla
a
DNAm: : LOCO
All test sets
R = 0.73, MAE = 0.45, P = 3.9e−42
b
1−NN: : LOCO
All test sets
R = 0.62, MAE = 0.72, P = 1.2e−27
c
DNAm: LOCO
Test set: Carnivora
R = 0.68, MAE = 0.45, P = 0.00033
d
1−NN: LOCO
Test set: Carnivora
R = 0.52, MAE = 0.95, P = 0.011
e
DNAm: LOCO
Test set: Primates
R = 0.51, MAE = 0.21, P = 0.0052
f
1−NN: LOCO
Test set: Primates
R = 0.37, MAE = 0.45, P = 0.049
g
DNAm: LOCO
Test set: Artiodactyla
R = 0.73, MAE = 0.5, P = 6.6e−09
h
1−NN: LOCO
Test set: Artiodactyla
R = 0.68, MAE = 0.28, P = 1.8e−07
Predicted (log) anAge lifespan
Observed (log) anAge lifespan
(de Magalhaes et. al., 2007)
Taxonomic Order
(1.*)Primates
(2.*)Proboscidea
(3.*)Lagomorpha
(4.*)Artiodactyla
(5.*)Carnivora
(6.*)Perissodactyla
(7.*)Cingulata
(8.*)Chiroptera
(9.*)Rodentia
(10.*)Didelphimorphia
(11.*)Diprotodontia
(12.*)Eulipotyphla
(13.*)Afrosoricida
(14.*)Sirenia
(15.*)Macroscelidea
(16.*)Tubulidentata
(17.*)Scandentia
(18.*)Dasyuromorphia
(19.*)Hyracoidea
(20.*)Pilosa
(21.*)Monotremata
(22.*)Dermoptera
(23.*)Microbiotheria
(25.*)Pholidota
(26.*)Paucituberculata

Figure 3.8: Taxonomic order breakdown of DNAm lifespan predictors and Phylogeny-based Predictors under LOCO. A breakdown of predictor performance in large taxonomic orders under LOCO. Panels **a** and **b** are identical to those of Figure 3.2c and Figure 3.2d. Panels **c**-**h** show large test set predictions. We compared prediction performance between DNAm elastic net predictors and 1-Nearest-Neighbor predictor (KNN). 1-Nearest-Neighbor predictor utilizes distances from the Mammalian phylogenetic TimeTree (Kumar et al., 2017). Panels **a**, DNAm predictor's test set predictions leave-one-clade-out (LOCO) training-test separation scheme; **b**, k-NN predictor's test set predictions under LOCO; **c**, **d**, DNAm and k-NN predictors, respectively, test set predictions of lifespan for all species belonging to Carnivora under LOCO; **e**, **f**, DNAm and k-NN predictors, respectively, test set predictions of lifespan for all species belonging to Primates under LOCO; **g**, **h** DNAm and k-NN predictors, respectively, test set predictions of lifespan for all species belonging to Artiodactyla under LOCO. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Numeric values can be found in Github repository shorvath/MammalianMethylationConsortium. Shaded areas represent 95% confidence intervals of the simple linear regression line.

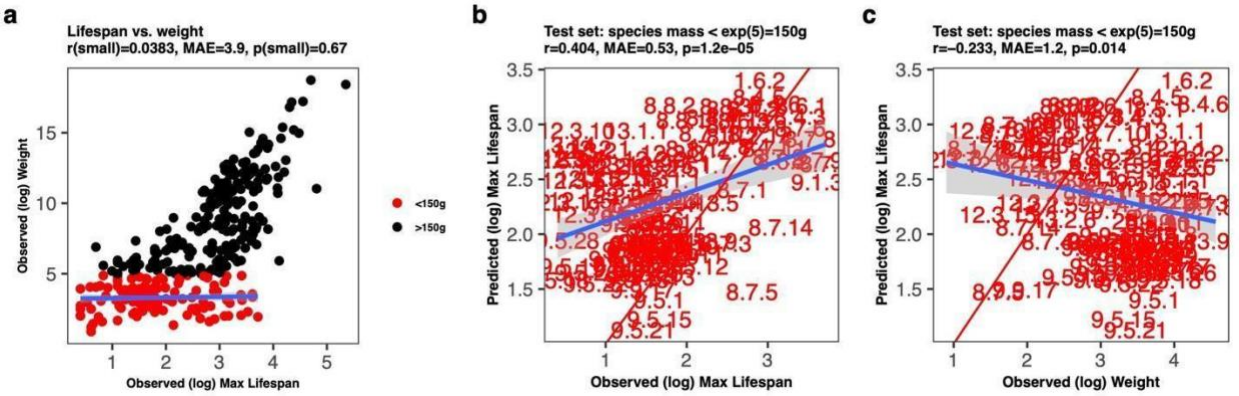

Figure 3.9: DNAm lifespan predictions on small-sized mammals. DNAm lifespan predictor trained on mammal species with an average weight over 150 grams (small mammals). Panels **a**, observed (log) adult body weight vs. observed (log) maximum lifespan in all mammalian species within the data set, color-coded by small-size indicator (more than 150 grams); **b**, test set predictions for the maximum lifespan in small-sized (<150 grams) mammalian species vs. observed (log) maximum lifespan; **c**, test set predictions for the maximum lifespan in small-sized (<150 grams) mammalian species vs. observed (log) adult body weight. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers are the mammalian species number annotation consistent with those of other figures. Numeric values can be found in Github repository shorvath/MammalianMethylationConsortium. Shaded areas represent 95% confidence intervals of the simple linear regression line.

3.4. Chapter acknowledgements

This chapter is a slightly modified version of my first-author paper published in the preprint, bioRxiv (C. Li et al., 2021), and has been reproduced here. Another version is being under review for a scientific journal.

CHAPTER 4

4. Marginal modeling for DNA methylation data

4.1. Methodology

4.1.1. Epigenome-wide association studies (EWAS)

Epigenome-wide association studies (EWAS) is the method of evaluating the marginal effect of each individual epigenetic biomarker to the outcome variable. It is usually implemented by regressing each CpG probe (See Section 1.1, a continuous covariate in the data set) to the outcome variable of interest, which, in the context of this data set, are life history traits such as species maximum lifespan, sexual maturation time, and gestation time. This constitutes a practice of multiple hypothesis testing, and is therefore often subject to multiple testing P-values corrections such as False Discovery Rate (FDR) (Y. Benjamini & Hochberg, 1995) and Bonferroni corrections (C. Bonferroni, 1936).

4.1.2. EWAS of life history traits

We restricted the EWAS of life history traits to 28,318 CpGs that were shown to work in two species of great importance in biomedical research: mice and humans. Toward this end, we used calibration/titration data (correlation with calibration exceeds 0.8) and mappability information as described in (Arneson et al.).

Since the distribution of maximum lifespan and other life history traits was highly skewed, we imposed a log-transformation on these phenotypes before conducting EWAS. We carried out four types of analyses that differ by how they deal with two potential confounders: adult weight and phylogeny. Our “generic” EWAS corresponds to a marginal correlation analysis where the

average methylation level of a given CpG per species was regressed on the (log-transformed) maximum lifespan using ordinary least squares regression. The second EWAS approach removed the confounding effect of average adult weight. To adjust for adult weight, we first regressed log maximum lifespan on log weight and formed residuals. Next the residuals become the dependent variables in the regression models. The third EWAS approach replaced ordinary least squares regression by phylogenetic regression, the variance-covariance matrix of which modeled evolutionary distances using branch lengths from the TimeTree project (Grafen, 1989; Kumar et al., 2017). The fourth EWAS approach adjusted for both average weight and for phylogenetic relationships. Due to the fact that phylogenetic regression takes into account sample covariance, it is more appropriate to report a phylogenetic independent contrast (PIC) as opposed to a simple scatter plot. Instead of using paired tip values from the tree, contrasts are calculated based on each node. The phylogenetic contrast model assumes that trait divergences occur independently at each node (Felsenstein, 1985).

We carried out EWAS analyses in the following tissues/organs for which a sufficient number of species (N>25 species) was available: skin (N = 137), blood (N = 133), liver (N = 147), skeletal muscle (N = 38), and brain (N = 26).

4.1.3. Functional enrichment algorithms for life history related cytosines

In order to make connections between identified significant CpG sites and meaningful biological pathways, one needs a robust statistical framework and large literature annotation database. One of the most widely used methods is hypergeometric (or sometimes Chi-square) enrichment test. We first select a certain number of top statistically significant genes. Second, we count the overlap between these genes and member genes in each pathway or literature collections of genes. These

counts combined with the overlap of array background genes are used to form a hypergeometric test. The purpose is to identify whether genes from a pathway are over-represented, accounting for the bias from the array design. However, EWAS identify significant CpGs instead of genes, and more than one CpG can be mapped to some genes. Thus we have decided to employ a genomic-region based enrichment method, using the R package for Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010b) in hg19 assembly. One major difference between GREAT and simple enrichment test is that GREAT uses genomic region overlaps instead of gene overlaps, more accurately accounting for CpG-to-gene mapping. The extension of gene regulatory regions was set at 50 kb and the other options were based on default settings. Since our EWAS focused on 28,318 CpGs that applied to both humans and mice, these probes were used as the background (Arneson et al., 2021). By specifying the background, GREAT analysis performed genomic-region based hypergeometric analysis, not confounded by gene sizes and uneven gene coverage.

In addition to gene set enrichment analyses, we conducted chromatin state enrichments using a universal annotation of the human genome annotation that is not specific to one cell or tissue type based on a stacked ChromHMM model recently generated based on over 1000 data sets from diverse human cell and tissue types (Vu & Ernst, 2020). ChromHMM is a multivariate hidden Markov model for characterizing and annotating patterns in histone marks, utilizing chromatin datasets such as ChIP-seq data (Ernst & Kellis, 2012, 2017). For each EWAS enrichment, we utilized a hyper-geometric test to assess significant overlap between chromatin states and the two sets of CpGs that are highly significant in either positive or negative correlations with maximum lifespan. The background set for these hyper-geometric enrichment tests were the 28,318 CpGs that mapped to both human and mouse.

4.1.4. Integrating human literature GWAS with mammalian EWAS

Our EWAS-GWAS based overlap analysis related the genomic regions found by our EWAS of maximum lifespans with the significant gene sets (top 2.5% of genes) found by published large-scale GWAS of various phenotypes, across body fat distribution, lipid panel outcomes, metabolic outcomes, neurological diseases, six DNAm based biomarkers, and other age-related traits (Ake T. Lu et al., 2021). A total of 102 GWAS traits were included in the enrichment database (**Supplementary Note 1**). This database includes six DNAm biomarkers based on four epigenetic age acceleration measures 1) pan-tissue epigenetic age adjusted for age-related blood cell counts, intrinsic epigenetic age acceleration (IEAA) (Steve Horvath, 2013; Horvath et al., 2016); 2) Hannum's blood-based DNAm age (Hannum et al., 2013); 3) DNAmPhenoAge (Levine et al., 2018); and 4) the mortality risk estimator DNAmGrimAge (Ake T Lu et al., 2019), as well as DNAm-based estimates of blood cell counts and plasminogen activator inhibitor 1 (PAI1) levels (Ake T Lu et al., 2019). For each GWAS trait, the MAGENTA software was used to calculate an overall GWAS P-value per gene. The P-values were calculated taken into account the most significant SNP association P-value within ± 50 kb of the gene adjusted for gene size, number of SNPs per kb, linkage disequilibrium, and other potential confounders (Segrè et al., 2010). The MAGENTA analysis was performed in MATLAB (2017 version). We restricted the analysis to genomic regions of GWAS genes present on the mammalian array. For each EWAS result, we studied the genomic regions from the top 500 CpGs per direction with strong associations with log-transformed life history traits (same thresholding described above, nominal $P < 3.5x10^{-7}$, Bonferroni corrected $P < 0.05$). To assess the overlap with a test trait, we selected the top 2.5% genes for each GWAS trait and calculated one-sided hypergeometric P-values based on genomic

regions. We report GWAS traits that led to a significant hypergeometric test (FDR corrected $P < 0.05$) for any EWAS of log-transformed life history traits. The number of background genomic regions in the hypergeometric test was based on the overlap between all genes in the GWAS and all genomic regions represented by the mammalian array. Enrichment p-values for the overlap between the genes implicated in EWAS and GWAS were based on genomic region-based hypergeometric tests as detailed in (C. Li et al., 2021).

4.2. Results

4.2.1. EWAS of maximum lifespan

We carried out epigenome-wide association studies (EWAS) to relate the methylation levels of individual CpGs to the various life history traits. To reduce biases resulting from different levels of sequence conservation, our EWAS of life history traits focused on n = 333 eutherian species, excluding Marsupial species. We performed four types of EWAS analyses adjusting for different confounders: (1) Lifespan; a direct regression analysis of lifespan (generic EWAS). (2) Weight-adjusted lifespan (AdjWeight); a regression analysis of maximum lifespan after adjustment for adult weight, which identifies lifespan-related CpGs that are independent of the body mass of the species. (3) Phylogenetic-adjusted lifespan (AdjPhylo); a phylogenetic regression model (de Magalhaes et al., 2007; Grafen, 1989) of lifespan, which adjusts for evolutionary relationships between species. (4) Phylogeny and Weight-adjusted lifespan (AdjPhyloWeight); a phylogenetic regression of lifespan after adjustment for average adult species weight. The results of these four categories of EWAS can be found in C. Li et al. (2021) tables. For brevity, we will focus on categories 1 and 2 since categories 3 and 4 led to qualitatively similar conclusions (C. Li et al., 2021).

Each analysis category is further subdivided by tissue type. The all-tissue analysis (denoted "All") ignored tissue type. Within species, mean methylation levels are highly correlated across tissue types (R > 0.95), but the all-tissue analysis may miss longevity mechanisms that are specific to tissues and organs. Therefore, we also present EWAS for five tissues for which there were a sufficiently large number of samples: blood (n = 141 species), skin (n = 146), liver (n = 151), muscle (n = 46) and brain (n = 34). We observed positive pairwise correlations between the all-tissue EWAS results and those of tissue-specific EWAS (Figure 4.3a-e**)**: such as blood (Pearson correlation R = 0.76), skin (R = 0.69), liver (R = 0.69), muscle (R = 0.49), and brain (R = 0.38). All tissue, Blood and skin lifespan EWAS are summarized in Figure 4.4, Figure 4.5, and Figure 4.6. To assess the robustness of maximum lifespan EWAS, we observed high agreements, in most tissues, between our generic EWAS (category 1) and a separate maximum lifespan EWAS using only samples obtained from animals that were younger than their species' average age of sexual maturity and younger than 5 (Figure 4.7).

We identified the genes that are proximal to CpGs that are statistically most correlated with maximum lifespan. These are as follows: lifespan was positively-correlated with a CpG in the distal intergenic region neighboring *TLE4* (Pearson R = 0.68, P = $2.9x10^{-46}$, Figure 4.2) and two CpGs near the promoter region of *HOXA4* (R = 0.67, Figure 4.2b and Figure 4.2c), and negatively-correlated with a CpG in an intron of *GATA3* (R = -0.65, P = $4.4x10^{-12}$, Figure 4.2d), exon in *ZBTB7B* (R = -6.1, P = $3.2x10^{-35}$, Figure 4.2e), and the promoter region of *C9orf106* (R = -0.6, P = $4.6x10^{-35}$, Figure 4.2f).

Many of these significant CpGs remain so after phylogenetic adjustment, such as the CpGs neighboring *TLE4, HOXA4*, *C9orf106, PKNOX2, LMX1B,* C15orf41, and *ZEB2* (P = $4.2x10^{-5}$, P = $4.8x10^{-3}$, P = $3.1x10^{-3}$, P = $3.6x10^{-4}$, P = $2.2x10^{-3}$, P = 4.8x10-7, P = $3.5x10^{-3}$, respectively, Figure

4.4) (C. Li et al., 2021). Phylogenetic EWAS (category 3 analysis) top CpGs are reported in the form of phylogenetic independent contrast (Figure 4.8), and the EWAS Z statistics agreements with generic lifespan EWAS (category 1 analysis) are summarized in Figure 4.9. Generic EWAS and phylogenetic EWAS Z statistics agreements are summarized in Figure 4.10.

All of the top-ranking CpGs mentioned above from the category 1 analysis remain in the top 500, in both directions, of weight adjusted EWAS (category 2 analysis) (Figure 4.4d), which indicates that these CpGs do not reflect confounding by body mass. But adjustment for adult weight (category 2) leads to a different set of top ranking CpGs: the top positively lifespan-related CpGs are in a promoter of *PKNOX2* (R = 5.4, P = 5.210^{-27}) and an intron of *LMX1B* (R = 0.51, P = $4.0x10^{-24}$) and the top CpGs negatively related to lifespan are in an intron of *C15orf41* (R = -0.55, P = $8.0x10^{-27}$) and an intron of *ZEB2* (R = -0.5, P = $5.1x10^{-23}$). Most of these top lifespan related CpGs in eutherians do not correlate with maximum lifespan across the 15 marsupial species (C. Li et al., 2021) which may be due to sequence differences or could reflect the low statistical power in marsupial species (only n=15 marsupial species).

Mammalian maximum lifespan is correlated with several other traits such as gestational time, and age at sexual maturity (Figure 4.11). Thus, we examined the degree of overlap between the EWAS of these evolutionary traits. At a Bonferroni corrected significance threshold of P = $1.8x10^{-6}$ (=0.05/28318), the methylation of 7429, 8218, and 5,962 CpGs were significantly associated with maximum lifespan, gestation time, and age at sexual maturity, respectively. An upset plot (generalization of Venn diagram) reveals that 329 CpGs relate significantly to all three life history traits (Figure 4.1b). Manhattan plots for EWAS of gestation time and age at sexual maturity are reported in Figure 4.12.

4.2.2. Gene set enrichment analysis of maximum lifespan

To uncover biological processes potentially linked to lifespan-related CpGs, we identified functional annotations associated with genes proximal to lifespan-related CpGs using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010a). GREAT automatically adjusts for biases arising from the array platform and biases of uneven coverage of genes.

The number of significant lifespan-related CpGs per tissue type depends on the underlying sample size (number of species). We imposed an upper limit of 500 on the number of significant CpGs and referred to the top 500 CpGs with a positive and negative correlation with lifespan as *lifespan.pos* set and *lifespan.neg* set, respectively. These CpGs are further subject to a Bonferroni corrected significance threshold (P < 1.8e-6) before the enrichment analysis. Detailed results can be found in Figure 4.1d, Figure 4.13, Figure 4.14, Figure 4.15, and C. Li et al. (2021).

CpGs that have a positive correlation with maximum lifespan implicate genes that play a critical role in development including the HOXL gene group (GREAT P = 1.2×10^{-5}, Figure 4.1d, Figure 4.13**)** based on the following genes *EVX1, HOXA2, HOXA3, HOXA4, HOXA5, HOXB1, HOXB2, HOXB3, HOXB4, HOXB7, HOXB8, HOXB9, HOXC4, HOXD10, HOXD8, HOXD9*) (C. Li et al., 2021). More significant enrichment for HOXL genes were obtained after adjusting the analysis for adult weight (reporting False Discovery Rate p-values as FDR) (GREAT FDR = 1.3×10^{-15}, Figure 4.1d) (C. Li et al., 2021). The EWAS of lifespan implicated embryonic organ morphogenesis with (category 2) or without (category 1) adjusting for adult weight (generic EWAS, GREAT FDR = 3.4×10^{-4}, Figure 4.13) (C. Li et al., 2021), weight-adjusted EWAS, GREAT FDR = 2.5×10^{-7}, Figure 4.13) (C. Li et al., 2021) and multicellular organism development (generic EWAS, GREAT FDR = 9×10^{-4}, Figure 4.1d, (C. Li et al., 2021) weight-adjusted EWAS,

GREAT FDR = $4.4x10^{-5}$, Figure 4.1d) (C. Li et al., 2021). Developmental pathways are even more enriched in skin samples, such as embryonic organ morphogenesis in generic EWAS (GREAT FDR = $2.7x10^{-30}$, Figure 4.15) (C. Li et al., 2021) and embryonic organ development in weight-adjusted EWAS (GREAT FDR = $1.4x10^{-15}$, Figure 4.15) (C. Li et al., 2021).

CpGs that are directly related to weight-adjusted maximum lifespan are enriched with genes involved in mouse phenotypes such as abnormal survival (GREAT P = $4.1x10^{-4}$) (C. Li et al., 2021) and mortality/aging (GREAT P = $7.2x10^{-5}$, Figure 4.13) (C. Li et al., 2021).

The GREAT enrichment analysis revealed that CpGs negatively related to lifespan are located next to genes that play a role in abnormal eye morphology according to mouse knockout studies (GREAT FDR = $2.3x10^{-4}$, Figure 4.1d) (C. Li et al., 2021), regulation of gene expression (GREAT FDR = $2.1x10^{-5}$, Figure 4.1d) (C. Li et al., 2021) and DNA-templated regulation of transcription (GREAT P=$4.0x10^{-5}$, Figure 4.1d) (C. Li et al., 2021).

Both negatively and positively lifespan related CpGs are located near genes that play different roles in mRNA processing (Figure 4.1d,Figure 4.14) (C. Li et al., 2021) and splicing including (*CELF1; CELF2; CELF6; DAZAP1; FAM172A; HNRNPA1; HNRNPK; HNRNPU ; JMJD6; MBNL1; MBNL2; NOVA2; QKI; RBFOX1; RBM15; RBM39; SF1; SON; SRPK1; SRPK2; SRSF12; TRA2A; TRA2B; YTHDC1*) (C. Li et al., 2021).

Our transcription factor analysis based on GREAT demonstrates that CpGs positively related to weight adjusted maximum lifespan are located near binding sites of transcription factors HOXA4 (GREAT FDR = $1.8x10^{-10}$, Figure 4.1d) (C. Li et al., 2021), GATA6 (GREAT FDR = $4.2x10^{-11}$), EVI1 (GREAT FDR = 0.001) while CpGs negatively related to lifespan are located near binding sites of transcription factor ER (GERAT FDR = $1.1x10^{-5}$ estrogen receptor) and IK3 (GREAT FDR = $2.7x10^{-4}$).

4.2.3. EWAS of cancer risk

Cancer risk has been extensively studied across mammalian species (Vincze et al., 2021). To obtain insights into cancer-related genes that contribute to epigenetic maximum lifespan, we regressed cytosines identified by EWAS of species-specific maximum lifespan on cancer risk as reported in the literature (Vincze et al., 2021). We designated the outcome variable of this analysis as cancer mortality risk (Vincze et al., 2021). The top hits are summarized in C. Li et al. (2021), and here we highlight 3 of the most significant CpG-neighboring genes, including tumor suppressor gene *CYLD* ($P = 4.4x10^{-7}$) (Fernández-Majada et al., 2016), oncogene *USP14* ($P = 1.7x10^{-6}$) (Zhu, Zhang, Gu, Li, & Wu, 2016), tumor suppressor *PRKAR2A* ($P = 1.3x10^{-6}$) (Saloustros et al., 2015), and a B-cell lymphoma tumor suppressor *PHIP* ($P = 2.3x10^{-6}$) (Weber et al., 2019). Interestingly, these CpGs remain largely significant after adjusting for average adult weight, placing *PHIP* as the top gene ($P = 1.4x10^{-6}$) (C. Li et al., 2021). None of these four cancer related CpGs overlap with CpGs from our EWAS of life history traits. Our EWAS of mammalian cancer risk implicates several genome-wide significant CpGs and neighboring genes that may serve as starting points for future studies of evolved cancer resistance. Another cancer outcome, cumulative incidence rate (Vincze et al., 2021), EWAS are reported in C. Li et al. (2021). This outcome variable covers less non-missing species. For concerns of statistical power, we focus on cancer mortality risk.

4.3. Additional figures

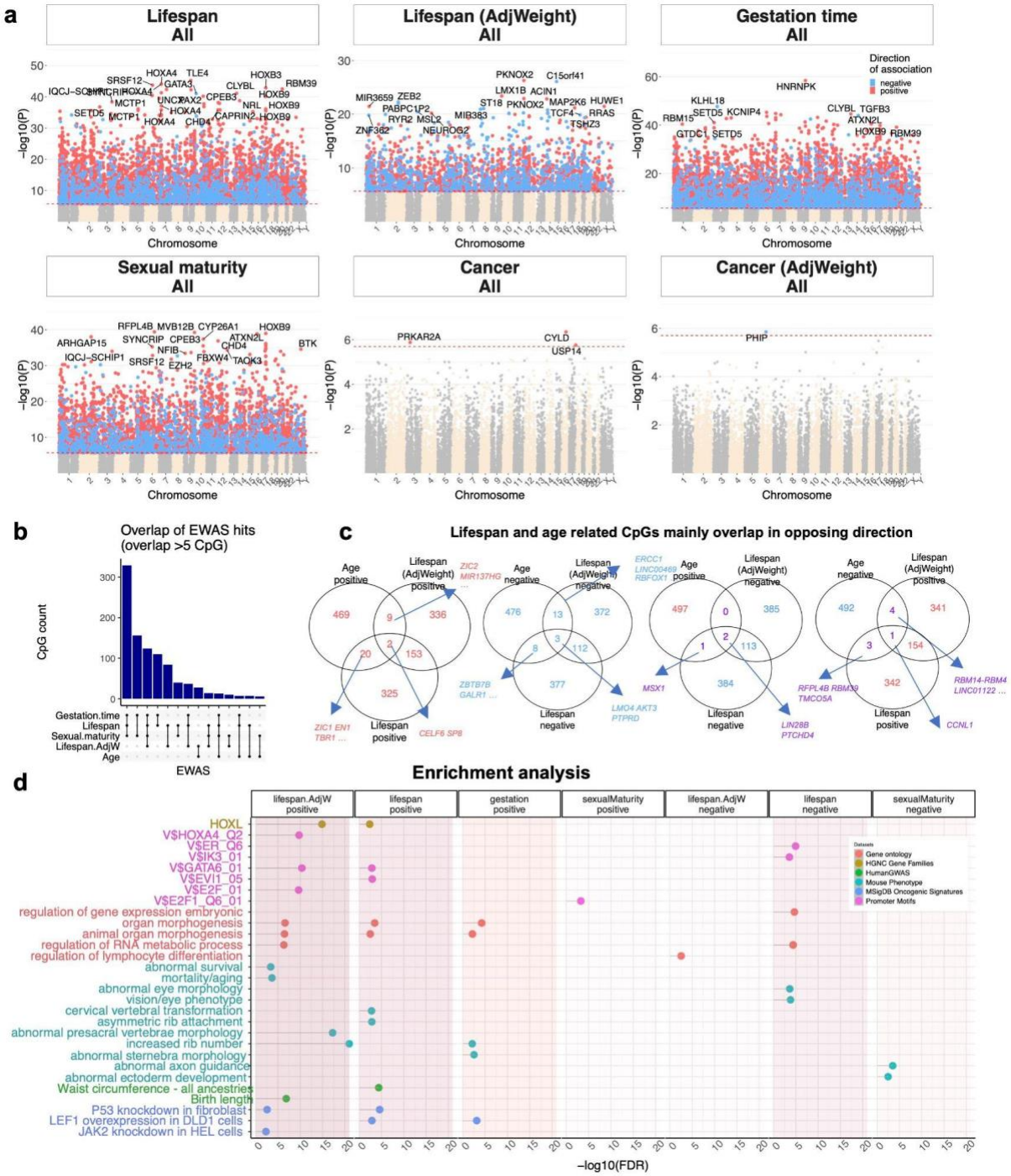

Figure 4.1: EWAS of eutherian log-transformed maximum lifespan, gestation time, age of sexual maturity, and risk of cancer. The figure represents the CpG specific association with maximum lifespan across n=333 eutherian species. All tissue samples were averaged by species. The associations with lifespan were examined with or without adjustment for adult weight of the species. a, Manhattan plots of EWAS results in 28,318 probes that were experimentally validated

to work in both mouse and human genomes. The coordinates are based on the alignment to the human hg19 genome. The red dotted line corresponds to a Bonferroni corrected two-sided p value $< 1.8x10^{-6}$. Individual CpGs with positive or negative correlations with maximum lifespan are colored in red and blue, respectively. The top significant CpGs are labeled by their respective neighboring genes. b, upset plot of the overlap in the top 1000 (500 per direction) significant CpGs for different EWAS models. c, Venn diagrams showing the overlap of CpGs associated with mammalian lifespan and the top 1000 CpGs that relate to chronological age in mammals (Ake T. Lu et al., 2021). Overlapping CpGs were labeled by neighboring genes. d, Gene set enrichment analysis of the genes proximal to CpGs associated with mammalian maximum lifespan, gestation time, and sexual maturity. We only report enrichment terms that are significant after adjustment for multiple comparisons (hypergeometric test false discovery rate <0.01) and contain at least five significant genes. The top two significant terms per enrichment database are shown in the panel.

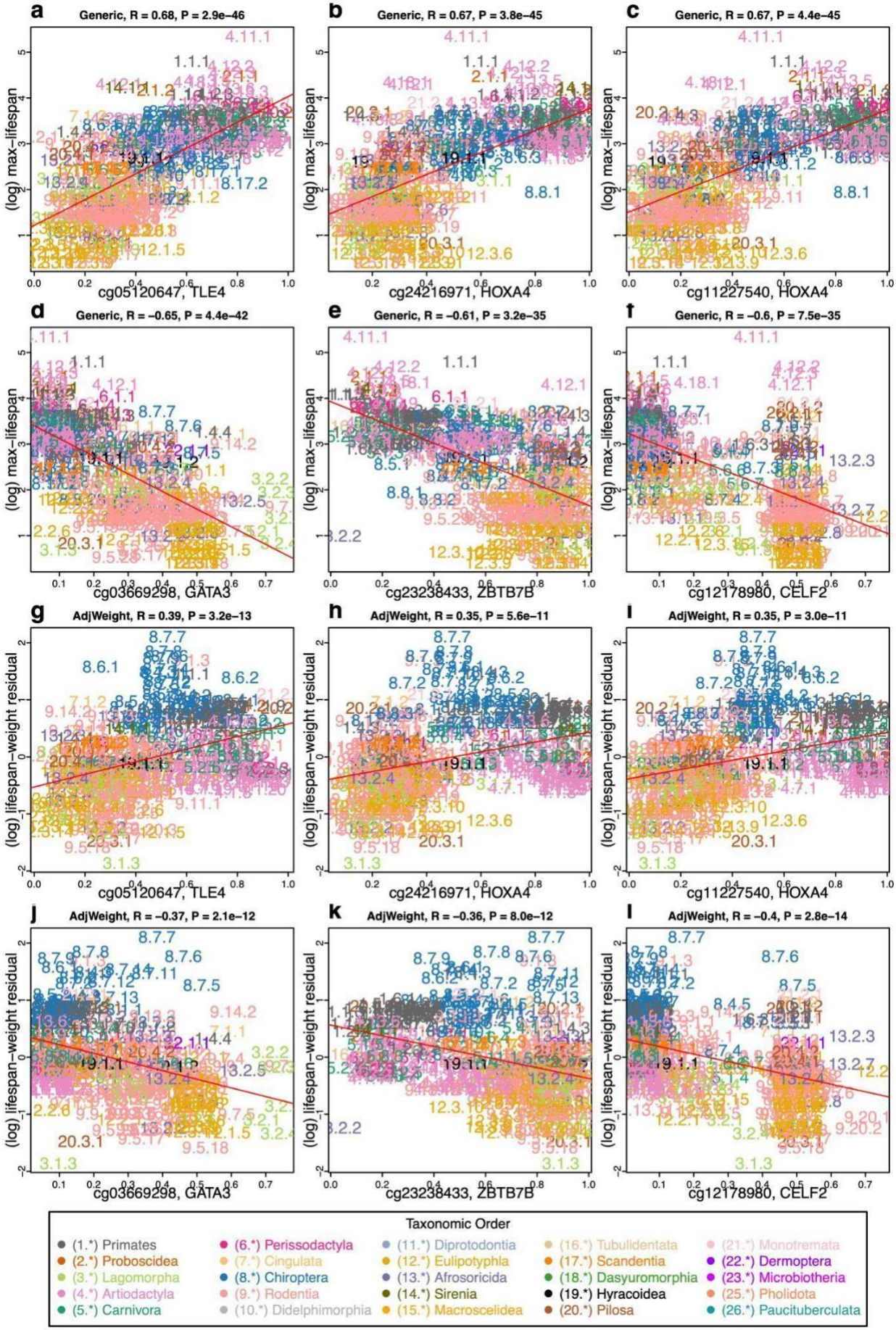
a
Generic, R = 0.68, P = 2.9e−46
b
Generic, R = 0.67, P = 3.8e−45
c
Generic, R = 0.67, P = 4.4e−45
d
Generic, R = −0.65, P = 4.4e−42
e
Generic, R = −0.61, P = 3.2e−35
f
Generic, R = −0.6, P = 7.5e−35
g
AdjWeight, R = 0.39, P = 3.2e−13
h
AdjWeight, R = 0.35, P = 5.6e−11
i
AdjWeight, R = 0.35, P = 3.0e−11
j
AdjWeight, R = −0.37, P = 2.1e−12
k
AdjWeight, R = −0.36, P = 8.0e−12
l
AdjWeight, R = −0.4, P = 2.8e−14
(log) max−lifespan
(log) lifespan−weight residual
cg05120647, TLE4
cg24216971, HOXA4
cg11227540, HOXA4
cg03669298, GATA3
cg23238433, ZBTB7B
cg12178980, CELF2
Taxonomic Order
(1.*) Primates
(2.*) Proboscidea
(3.*) Lagomorpha
(4.*) Artiodactyla
(5.*) Carnivora
(6.*) Perissodactyla
(7.*) Cingulata
(8.*) Chiroptera
(9.*) Rodentia
(10.*) Didelphimorphia
(11.*) Diprotodontia
(12.*) Eulipotyphla
(13.*) Afrosoricida
(14.*) Sirenia
(15.*) Macroscelidea
(16.*) Tubulidentata
(17.*) Scandentia
(18.*) Dasyuromorphia
(19.*) Hyracoidea
(20.*) Pilosa
(21.*) Monotremata
(22.*) Dermoptera
(23.*) Microbiotheria
(25.*) Pholidota
(26.*) Paucituberculata

Figure 4.2: Top CpGs related to log-transformed maximum lifespan in eutherians.Scatter plots of CpG methylation level (x-axis) versus log-transformed maximum lifespan (y-axis) for **a**, **b**, **c** the top three positively-correlated CpGs and **d**, **e**, **f** the top three negatively-correlated CpGs. **g**–**l**. Corresponding scatter plots to **a**–**f** for weight-adjusted maximum lifespan. The y-axis reports the residuals resulting from regressing log-transformed maximum lifespan on log-transformed adult weight. Each observation corresponds to one of 333 different eutherian species and is colored and labeled by mammalian number as in Figure 3.1. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Red solid line represents the perfect prediction line.

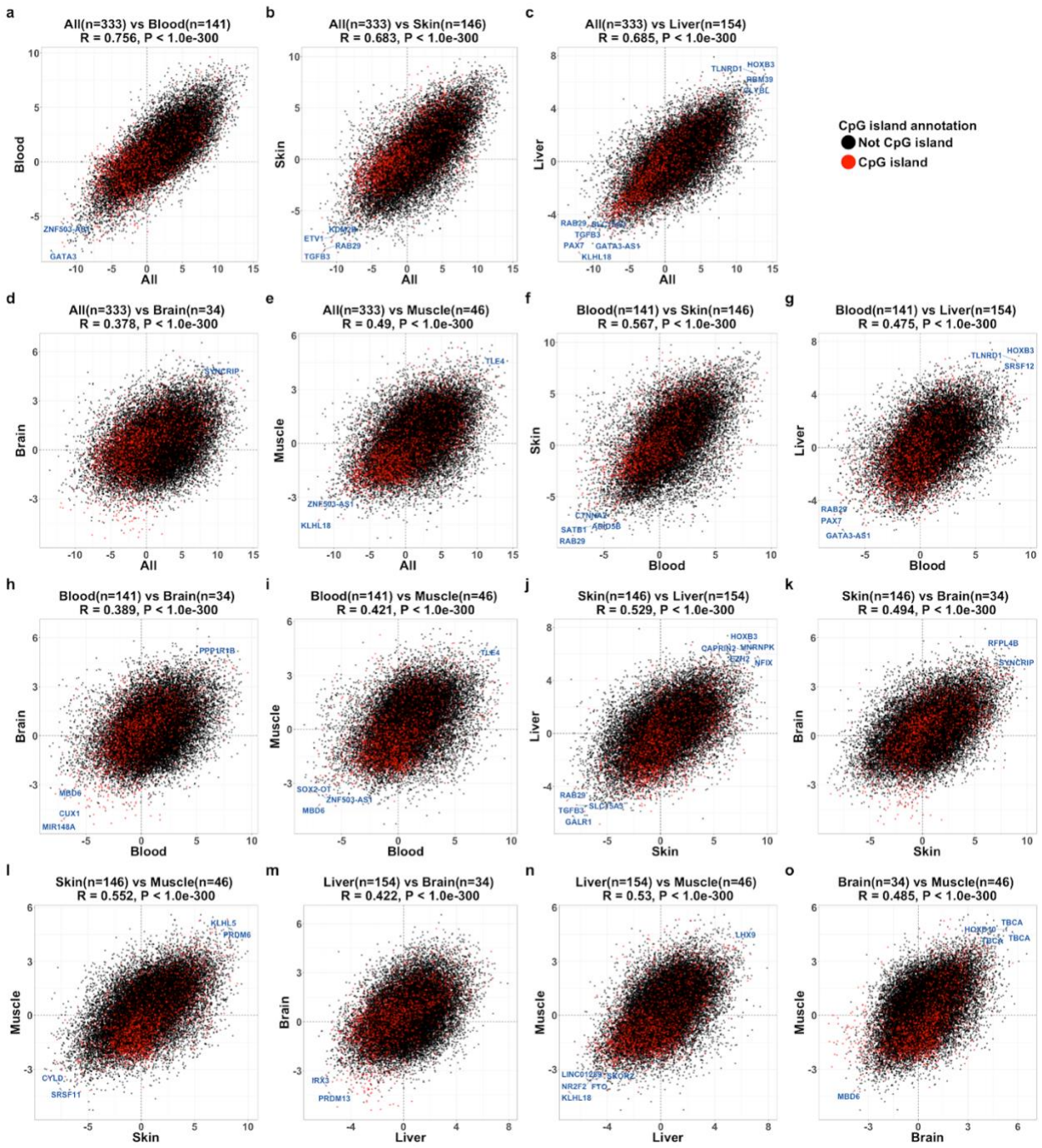

Figure 4.3: Generic Lifespan EWAS in different tissues from Eutherian species. Scatter plot of CpG Z statistics agreements between tissues, color-coded by human CpG island annotations (not island: black, island: red). Both x- and y-axes are CpG Z statistics for the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). Panels show agreements between **a,** blood vs. all, **b,** skin vs. all, **c,** liver vs. all, **d,** brain vs. all, **e,** muscle vs. all, **f,** skin vs. blood, **g,** liver vs. blood, **h,** brain

vs. blood, **i**, muscle vs. blood, **j,** liver vs. skin, **k,** brain vs. skin, **l,** muscle vs. skin, **m,** brain vs. liver, **n,** muscle vs. liver, **o,** muscle vs. brain. Panel titles report r and p as Pearson's correlation and p-values, respectively.

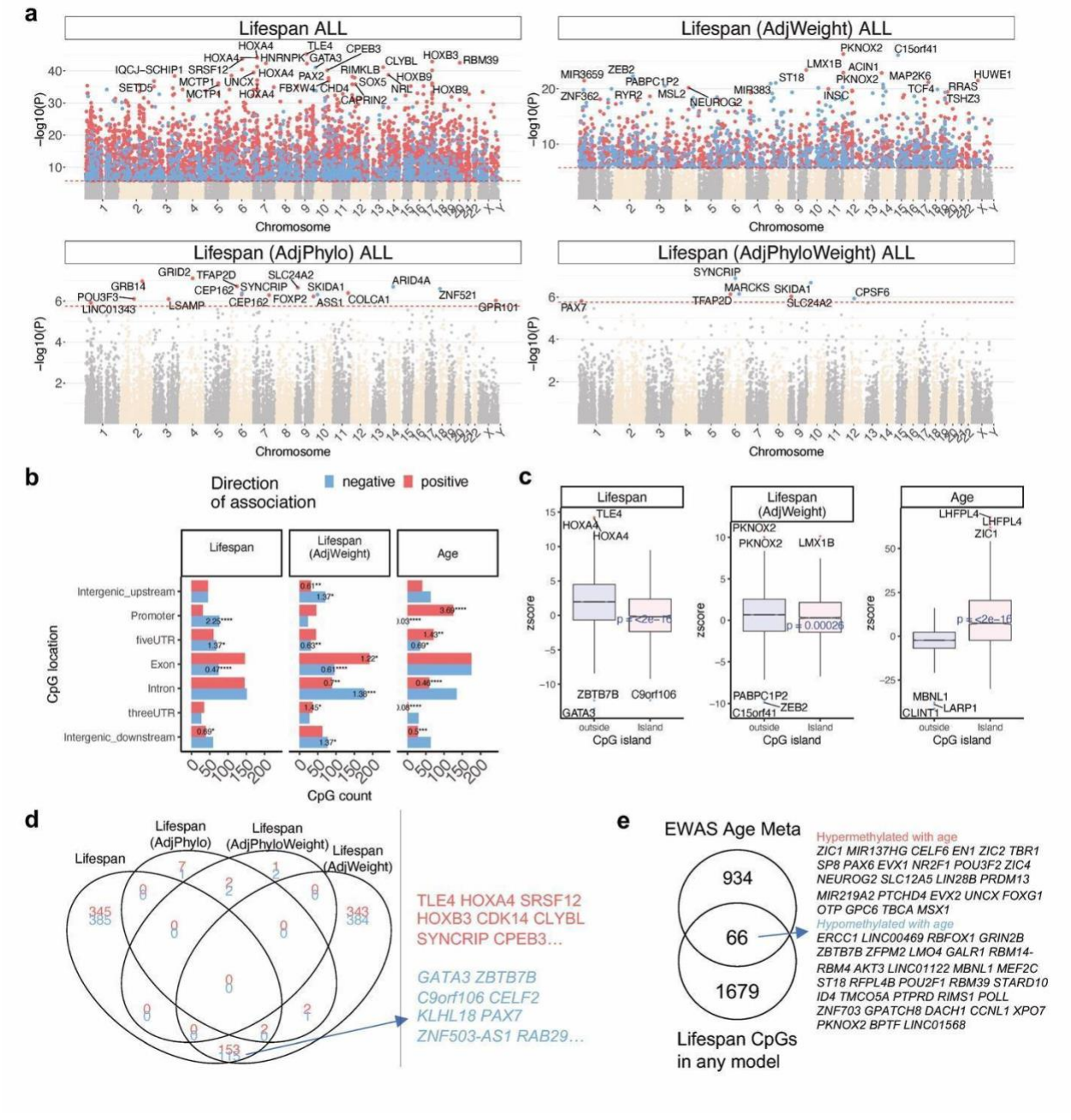

Figure 4.4: EWAS of significant CpGs related to mammalian maximum lifespan, adjusted by weight and phylogeny. Panel **a** are Manhattan plots reporting Manhattan plots of lifespan, lifespan EWAS adjusted by weight (AdjWeight), lifespan EWAS adjusted by phylogeny (AdjPhylo), and lifespan adjusted by both weight and phylogeny (AdjPhyloWeight). The background probes were limited to the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). **b**, Location of the top CpGs in each tissue relative to the closest transcriptional start site. A panel for the top 1000 age related CpGs was added to the figure for comparison (Ake T. Lu et al., 2021). The changes in gene regions were tested by a hypergeometric test in proportion to the background. The odd ratios and p-values (* <0.05, **<0.01, ***<0.001, ****<0.0001) of changes are reported for each bar. **c**,

Boxplot of association with mammalian maximum lifespan by human CpG island status. The mean difference was tested by Student T-test. **d**, Venn diagram of the overlap in the top 1000 (500 per direction) significant CpGs for different models of EWAS of lifespan from panel **a**. The overlap hits were labeled by neighboring genes. **e**, Overlap of CpGs associated with mammalian lifespan and the top 1000 CpGs that relate to chronological age in mammals (Ake T. Lu et al., 2021). Blood and skin specific results are reported in Figure 4.5, Figure 4.6, and Figure 4.7.

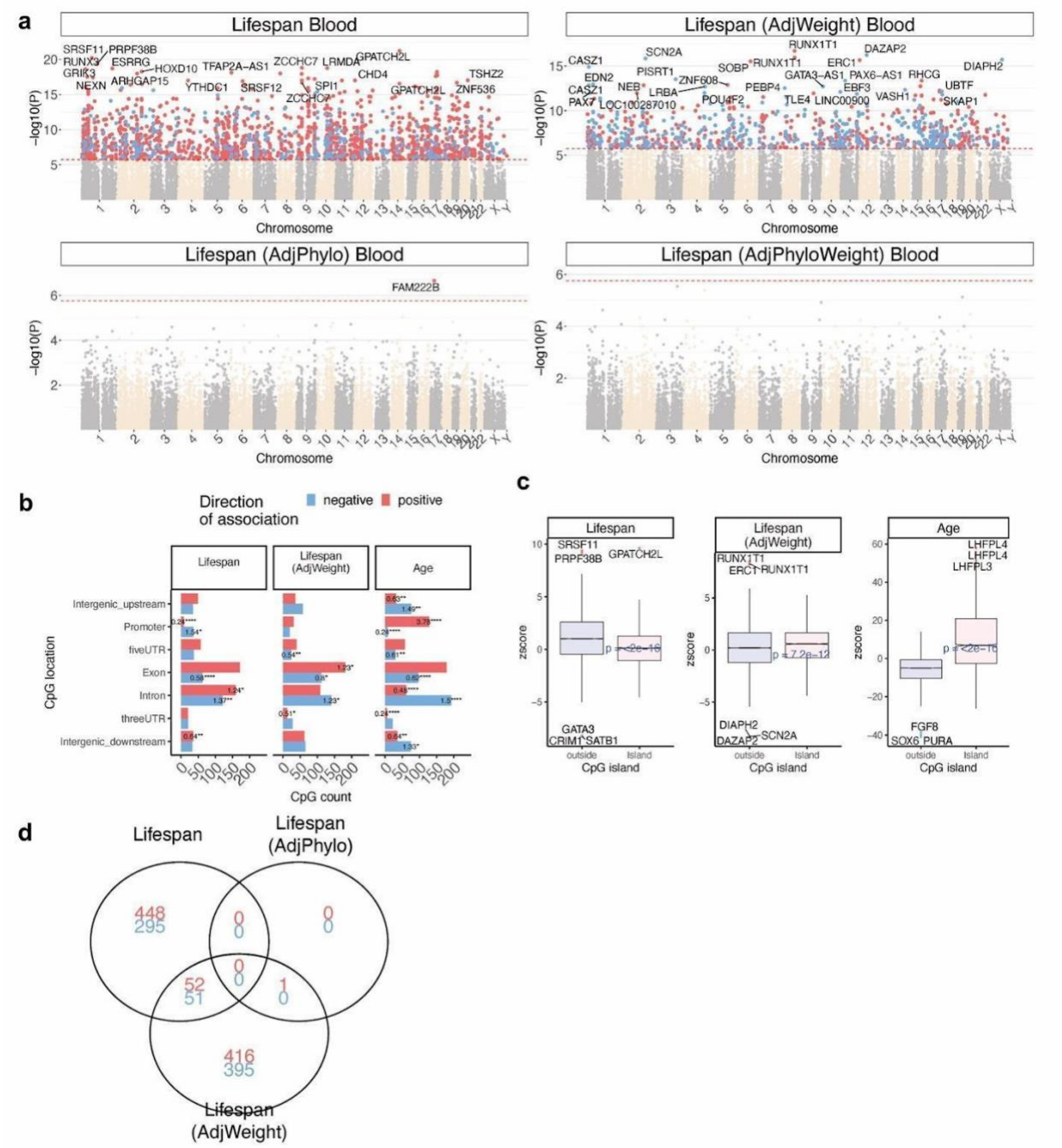

Figure 4.5: EWAS of mammalian maximum lifespan in blood. The associations were examined with four different models: 1) lifespan: each species as a datapoint in the model regardless of evolutionary distance. 2) lifespan adjusted for average species weight. 3) lifespan adjusted for evolutionary distance by phylogenetic regression. The evolutionary tree was acquired from TimeTree database. 4) lifespan adjusted for both average adult species weight and evolutionary distance. Panel **a,** Manhattan plots (Kumar et al., 2017) of EWAS of maximum lifespan in the set

of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). The coordinates are based on the alignment to the human hg19 genome. The direction of associations with $p < 0.001$ (red dotted line) is highlighted by red (hypermethylated) and blue (hypomethylated) colors. Some top CpGs were labeled by the neighboring genes, **b,** Location of top CpGs relative to the closest transcriptional start site. A panel for the top 500 age-related CpGs in each direction was added to the figure for comparison (Ake T. Lu et al., 2021). The changes in each gene region was tested by Fisher's exact test based on the same background. The odds ratios and p-values (* <0.05, **<0.01, ***<0.001, ****<0.0001) of changes are reported for each bar. **c,** Boxplot of association with mammalian maximum lifespan by human CpG island status. The mean difference was tested by a Student's T test. A panel for the top 1000 age-related CpGs was added to the figure for comparison, **d,** Venn diagram of the overlap in the top 1000 (500 per direction) significant CpGs for different models of EWAS of lifespan. The Venn diagram does not show AdjPhyloWeight because it contains zero CpG probe past the significance threshold.

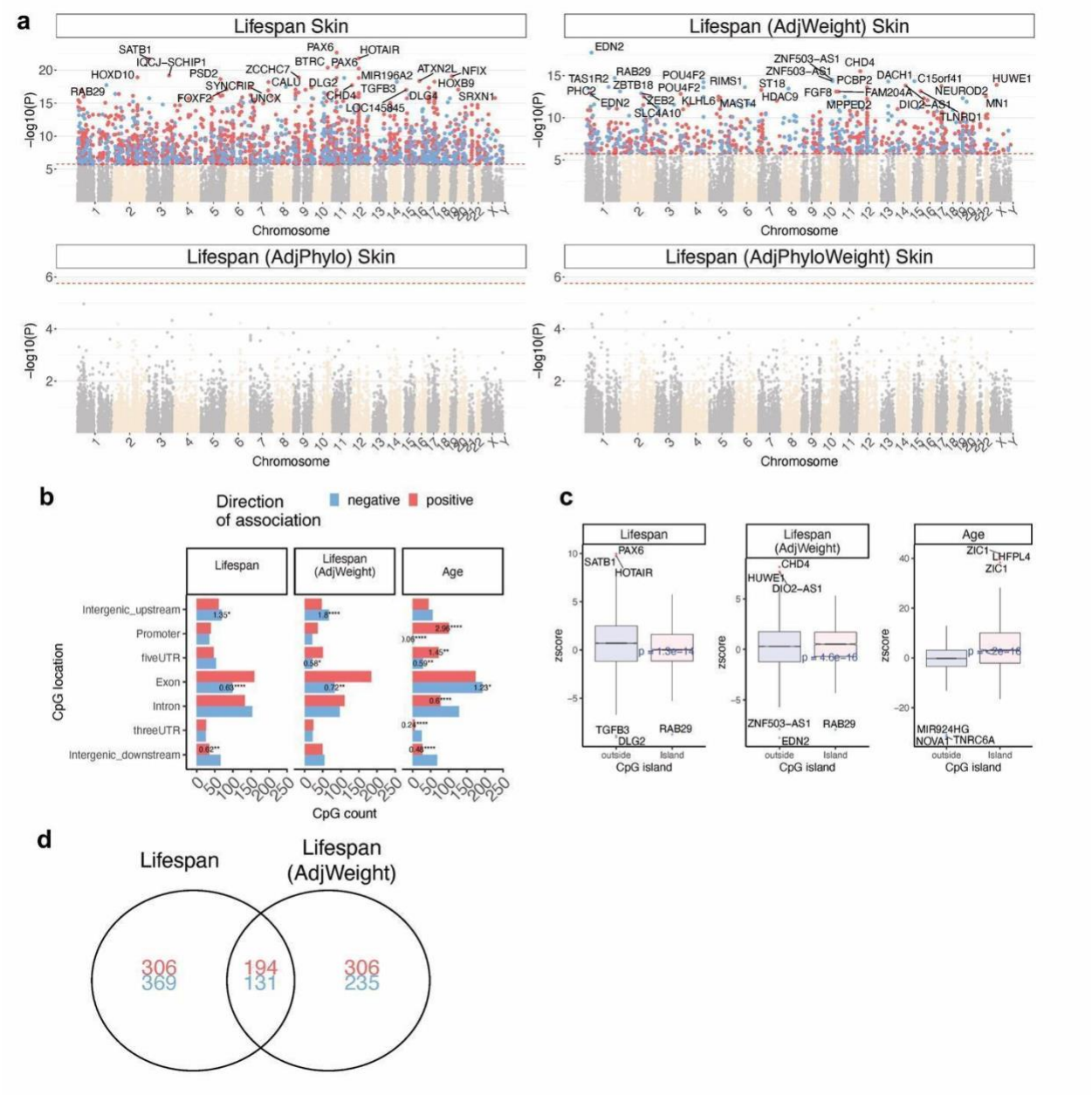

Figure 4.6: EWAS of mammalian maximum lifespan in skin. The associations were examined with four different models: 1) lifespan: each species as a datapoint in the model regardless of evolutionary distance. 2) lifespan adjusted for average species weight. 3) lifespan adjusted for evolutionary distance by phylogenetic regression. The evolutionary tree was acquired from TimeTree database (Kumar et al., 2017). 4) lifespan adjusted for both average adult species weight and evolutionary distance. Panel **a,** Manhattan plots of EWAS of maximum lifespan in the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). The coordinates are based on the alignment to the Human hg19 genome. The direction of associations with $p < 0.001$ (red dotted line) is highlighted by red (hypermethylated) and blue (hypomethylated) colors. The top few CpGs were labeled by the neighboring genes, **b,** Location of top CpGs in each tissue relative to the closest

transcriptional start site. A panel for the top 1000 age-related CpGs was added to the figure for comparison. The changes in each gene region were tested by Fisher's exact test based on the same background. The odds ratios and p-values (* <0.05, **<0.01, ***<0.001, ****<0.0001) of changes are reported for each bar. **c,** Boxplot of association with mammalian maximum lifespan by human CpG island status. The mean difference was tested by a student's T test. A panel for the top 1000 age-related CpGs was added to the figure for comparison, **d** Venn diagram of the overlap in the top 1000 (500 per direction) significant CpGs for different models of EWAS of lifespan.

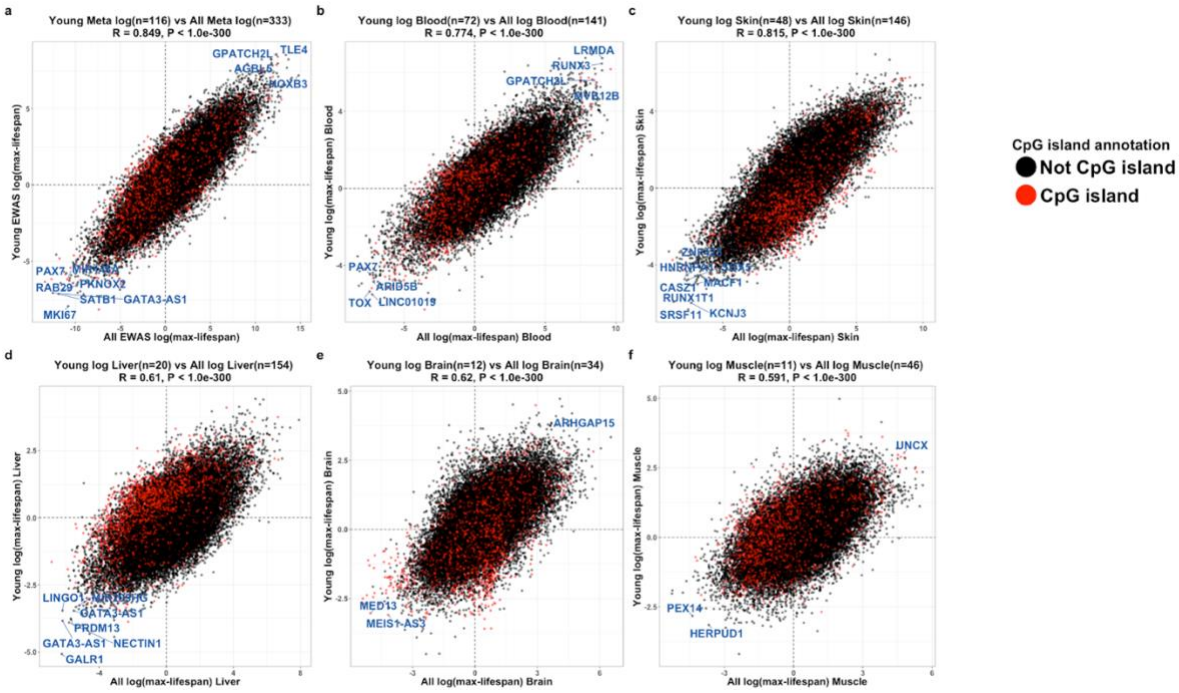

Figure 4.7: Generic EWAS agreements between all samples and young samples. Agreements between EWAS based on young samples and EWAS based on all available samples. Young samples are defined as samples younger than five years of age and before the age of sexual maturity. Panels show agreements between, **a** all tissue all vs. young generic EWAS, **b,** all vs. young generic EWAS in blood, **c,** all vs. young generic EWAS in skin, **d,** all vs. young generic EWAS in liver, **e,** all vs. young generic EWAS in brain, **f,** all vs. young generic EWAS in muscle. Panel titles report r and p as Pearson's correlation and p-values, respectively.

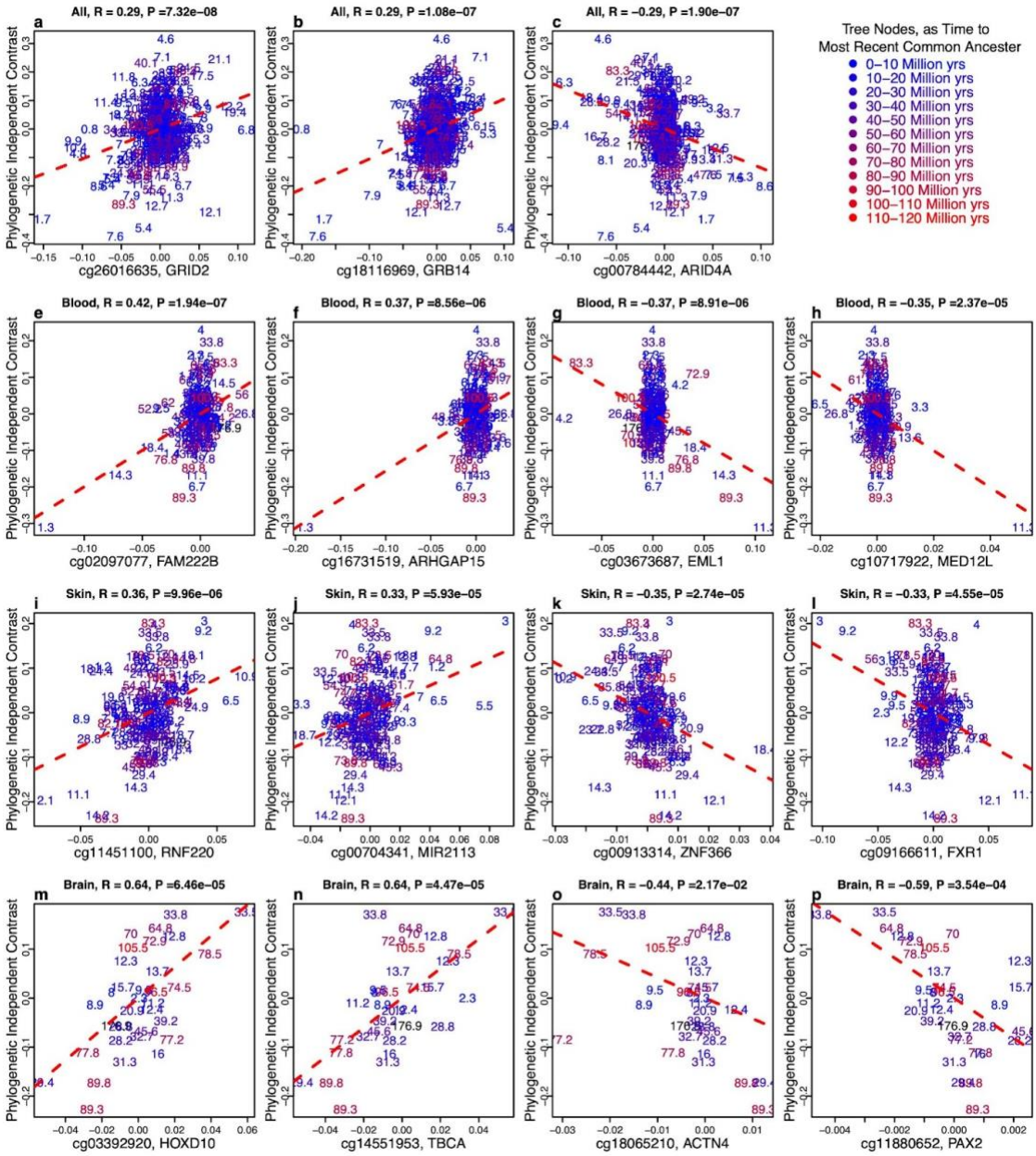

Figure 4.8: Top Significant CpG sites in a phylogenetic independent contrast plot, Eutherians. Scatter plot of CpG methylation and maximum lifespan, transformed and scaled to phylogenetic independent contrasts, based on all available samples. In order to properly visualize sample correlations, phylogenetic independent contrast plots select parent nodes that are of relatively similar distances to each other (Felsenstein, 1985). We color-coded these common ancestor nodes as time to present, in millions of years. Panels show scatter plots of top three CpGs from **a–c,** all tissues, **b–g,** top four CpG from blood tissues, **h–k,** top four CpGs from skin tissues, **l–o,** top four CpGs from brain tissues. P-values reported are based on phylogenetic generalized least squared (GLS) regression. Panel titles report r and p as Pearson's correlation and p-values, respectively.

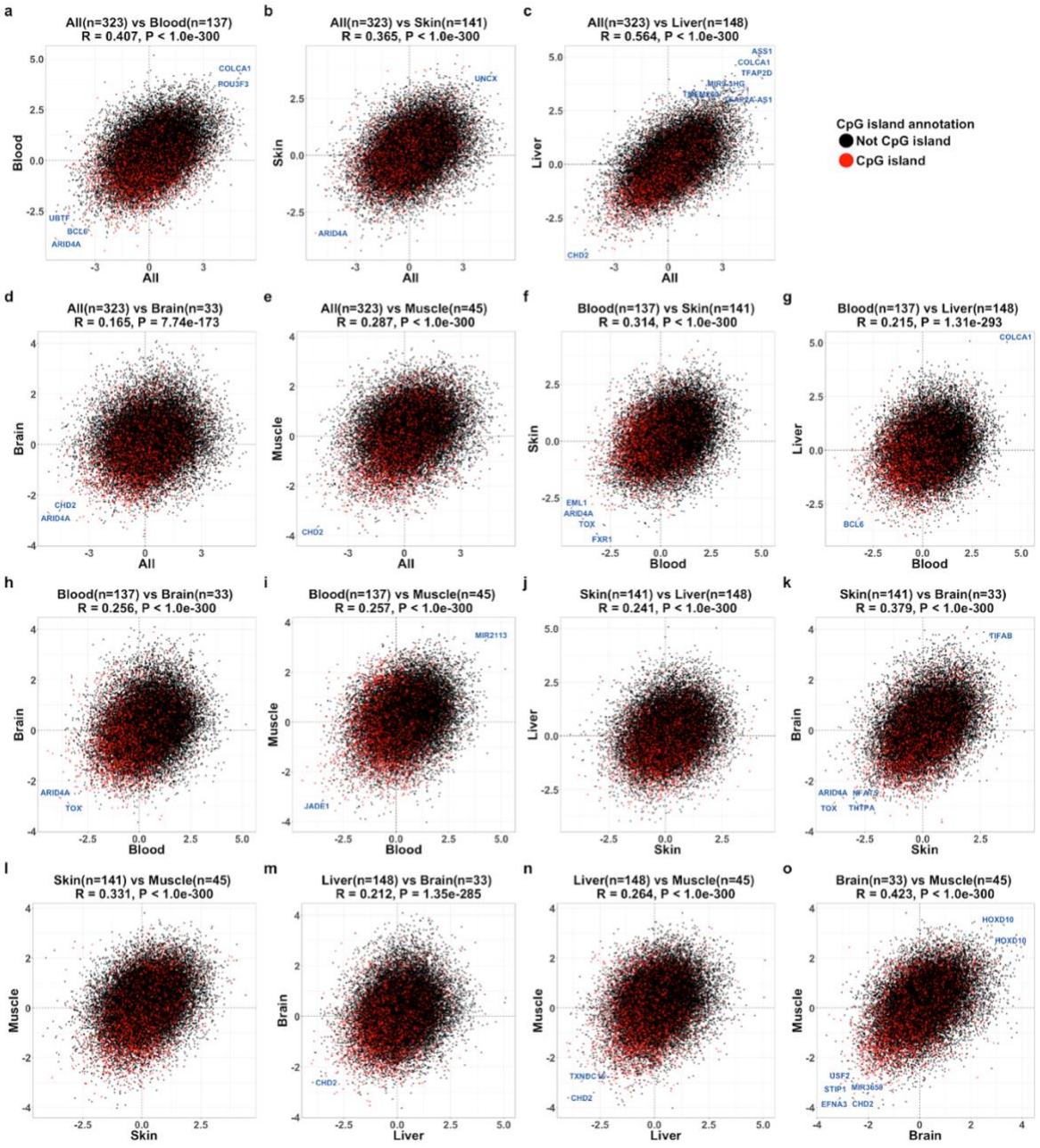

Figure 4.9: Phylogenetic EWAS agreement in various tissues, Eutherians. Scatter plot of CpG Z statistics between tissues, color-coded by human CpG island annotations (not island: black, island: red). Both x- and y-axes are CpG Z statistics for the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). Panels show agreements between **a,** blood vs. all, **b,** skin vs. all, **c,** liver vs. all, **d,** brain vs. all, **e,** muscle vs. all, **f,** skin vs. blood, **g,** liver vs. blood, **h,** brain vs. blood, **i**, muscle vs. blood, **j,** liver vs. skin, **k,** brain vs. skin, **l,** muscle vs. skin, **m,** brain vs. liver, **n,** muscle vs. liver, **o,** muscle vs. brain. Panel titles report r and p as Pearson's correlation and p-values, respectively.

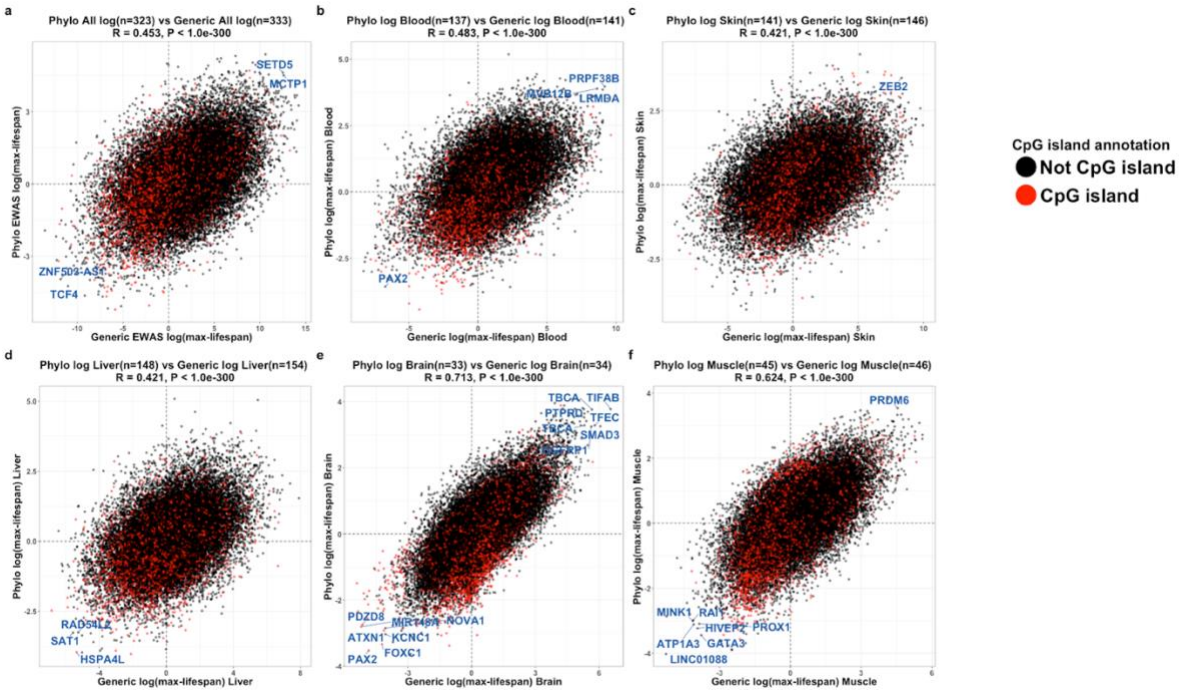

Figure 4.10: Simple linear regression (generic) and phylogenetic regression EWAS agreement. Scatter plot of CpG Z statistics across phylogenetic Generic EWAS vs. Phylogenetic EWAS. Similar to Figure 4.7, panel titles and axes labels report agreements between EWAS analyses. Panels show agreements between, **a** all tissue phylogenetic vs. generic EWAS, **b,** phylogenetic vs. generic EWAS in blood, **c,** phylogenetic vs. generic EWAS in skin, **d,** phylogenetic vs. generic EWAS in liver, **e,** phylogenetic vs. generic EWAS in brain, **f,** phylogenetic vs. generic EWAS in muscle. Panel titles report r and p as Pearson's correlation and p-values, respectively.

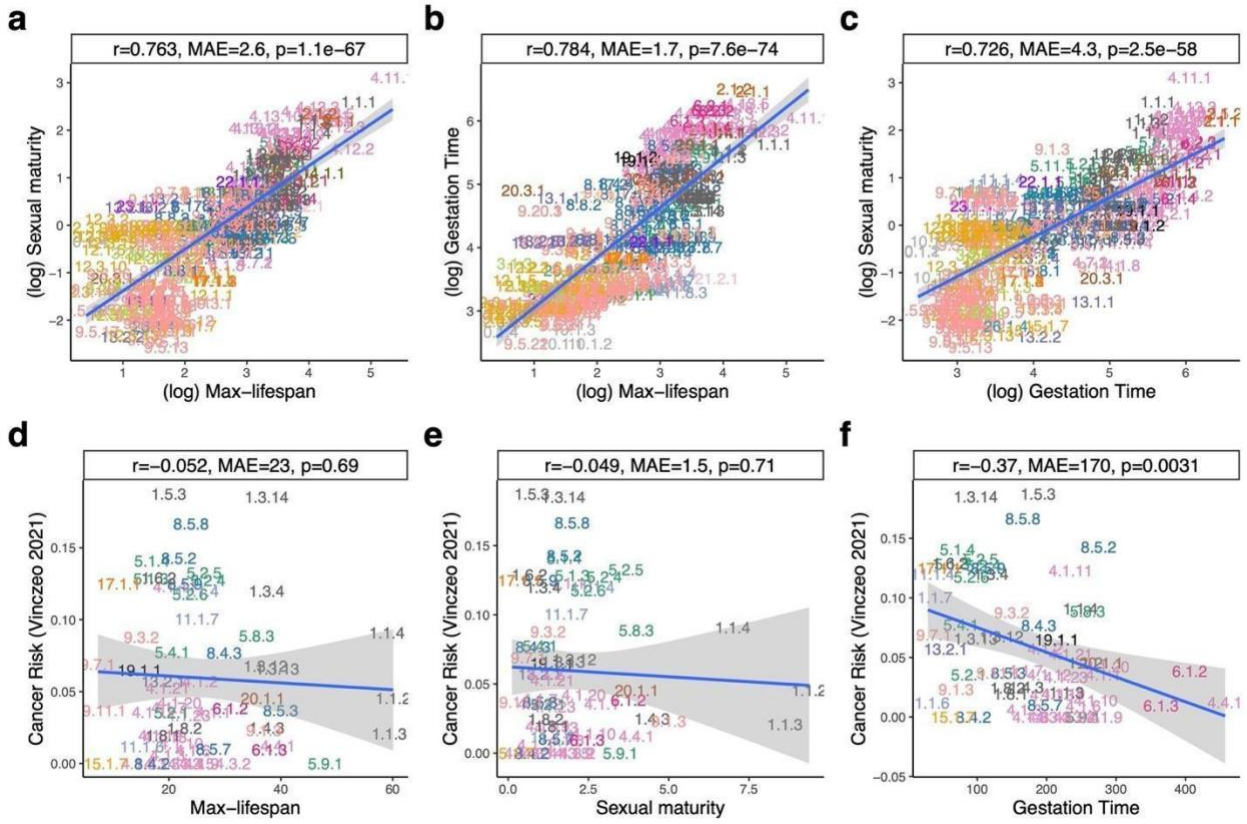

Figure 4.11: Mammalian life history traits relations. Panels show log-transformed relationships between observed variables of **a**, age at sexual maturity and maximum lifespan, **b**, gestation time and maximum lifespan, **c**, sexual maturity time and gestation time, **d**, cancer risk and maximum lifespan, **e**, cancer risk and sexual maturity, **f**, cancer risk and gestation time. MAE abbreviates median absolute errors from the regression errors; r and p are Pearson's correlation and p-values, respectively. Numbers and colors are the mammalian species number and order annotation consistent with those of other figures. Shaded areas represent 95% confidence intervals of the simple linear regression line.

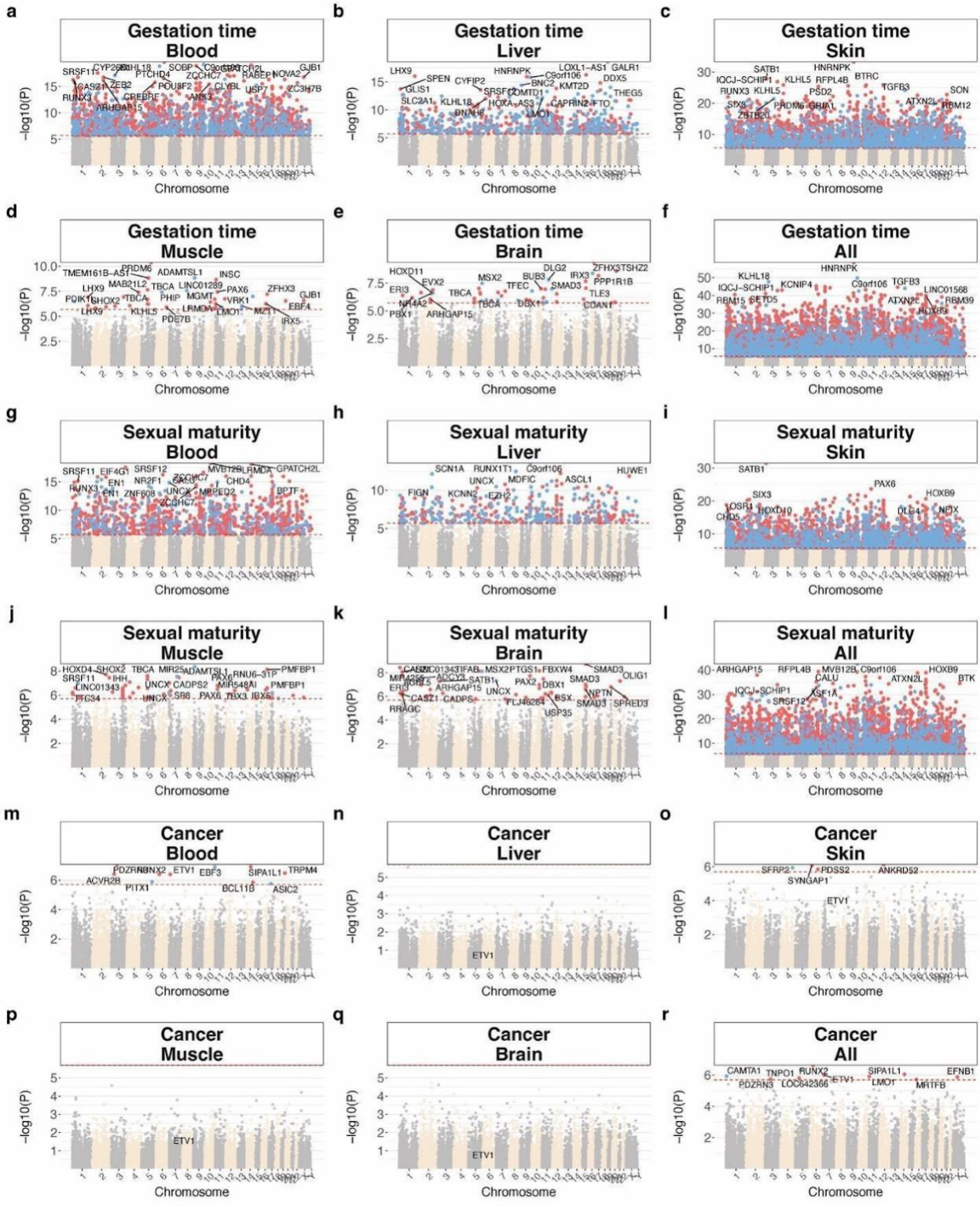
a
Gestation time
Blood
b
Gestation time
Liver
c
Gestation time
Skin
d
Gestation time
Muscle
e
Gestation time
Brain
f
Gestation time
All
g
Sexual maturity
Blood
h
Sexual maturity
Liver
i
Sexual maturity
Skin
j
Sexual maturity
Muscle
k
Sexual maturity
Brain
l
Sexual maturity
All
m
Cancer
Blood
n
Cancer
Liver
o
Cancer
Skin
p
Cancer
Muscle
q
Cancer
Brain
r
Cancer
All
−log10(P)
Chromosome

Figure 4.12: EWAS of significant CpGs related to mammalian life history traits, maximum lifespan, gestation time, sexual maturity time, and cancer risk. Manhattan plots of tissue-specific generic EWAS results for gestation, age at sexual maturity, and cancer risk. Red dotted line represents our Bonferroni-adjusted significance level. Manhattan plots report the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8).

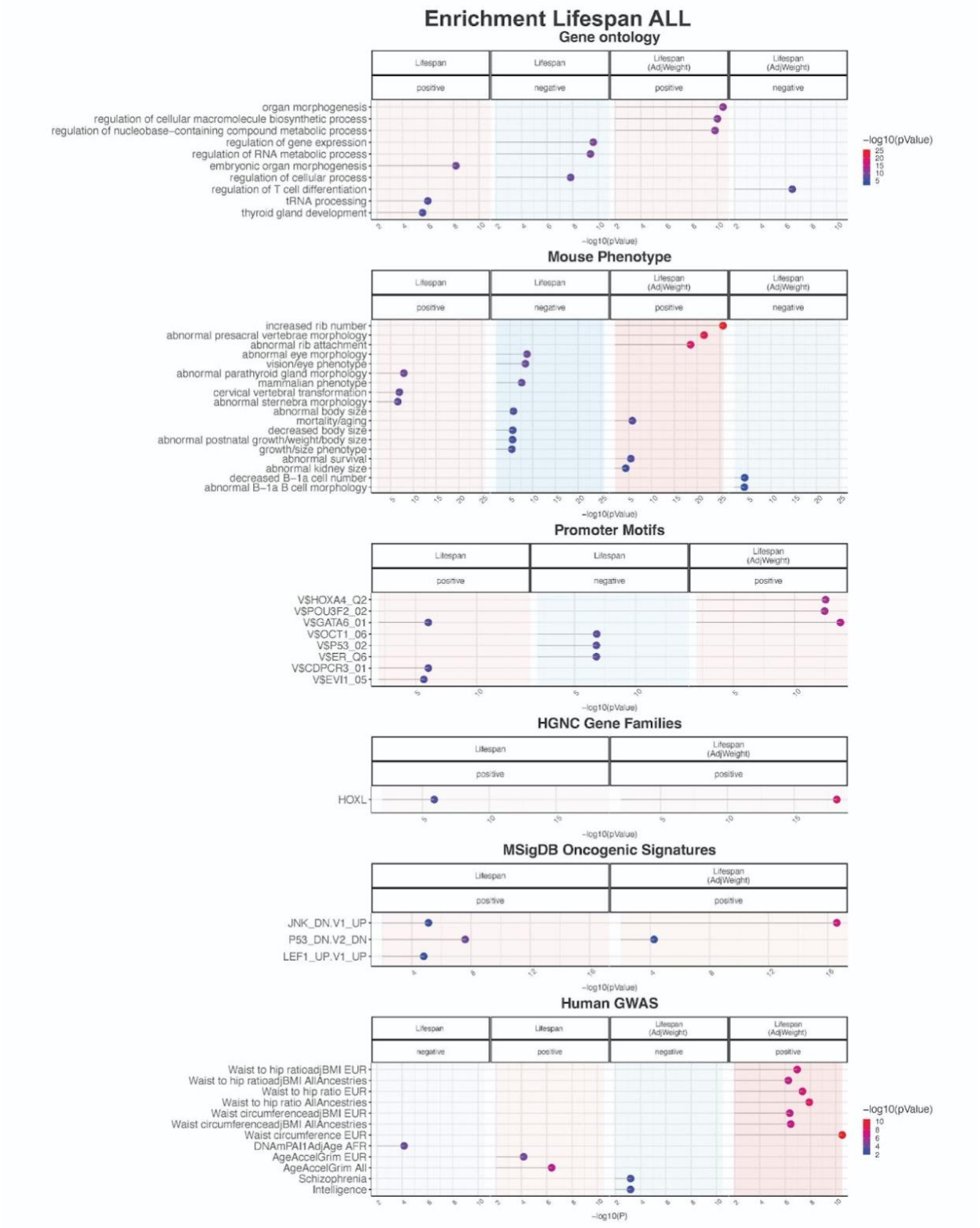

Figure 4.13: Gene set enrichment analysis of significant CpGs related to mammalian maximum lifespan. The gene-level enrichment was done using GREAT analysis using human background. Foreground selection is consistent with the description in the methods section. The background

probes were limited to the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). Human GWAS enrichment was calculated by a hypergeometric test of the top 2.5% genes involved in GWAS of complex traits-associated genes with the top lifespan-related gene regions in our analysis. The biological processes were reduced to parent ontology terms using the "rrvgo" package (Method). Input: Lifespan negative/positive, 500/500 CpGs; Lifespan (AdjWeight) negative/positive, 500/500. In each panel, the columns with no significant terms were removed to simplify the figure. Panels only show entries below a p-value threshold of $p<1x10-4$.

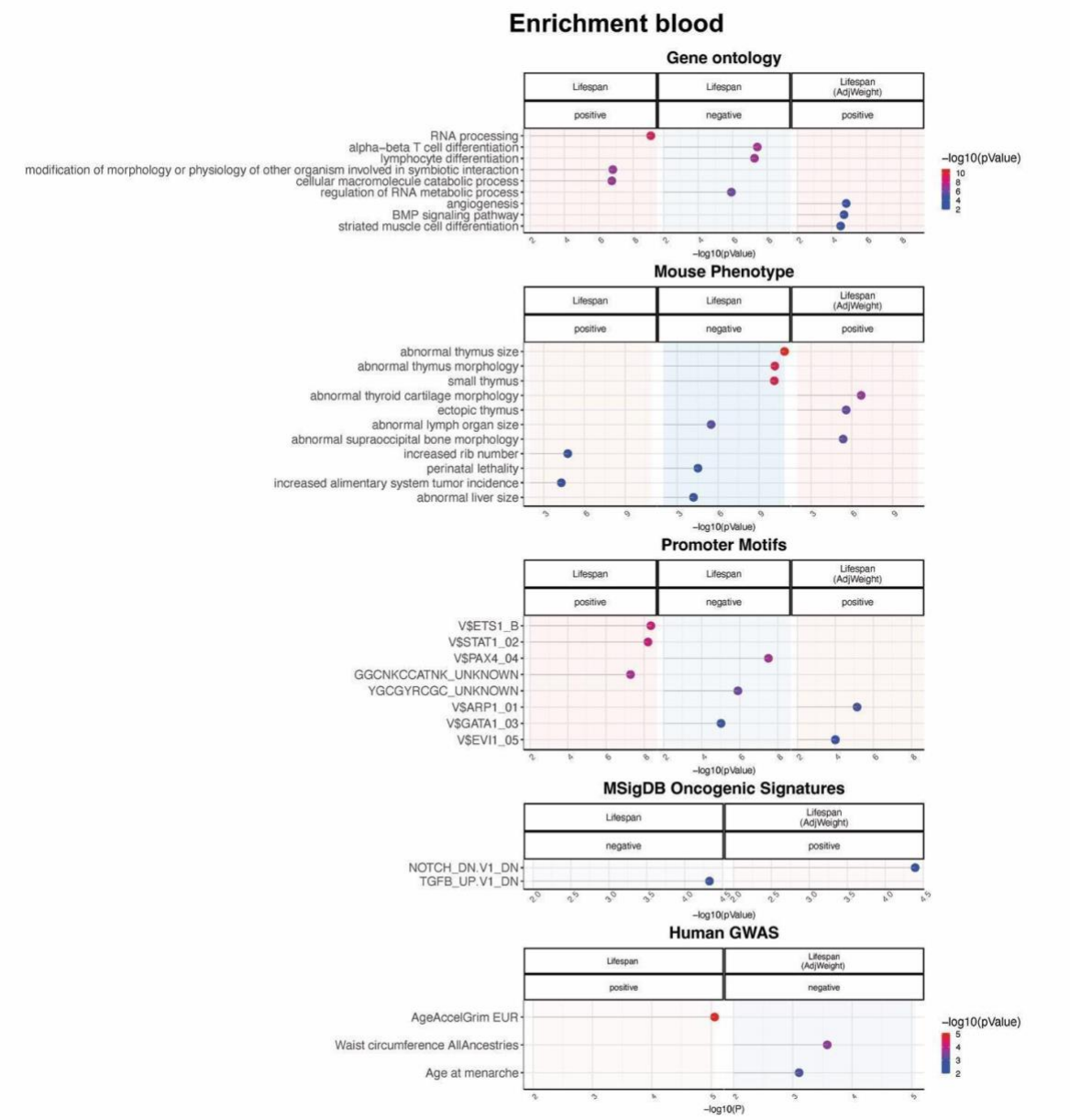

Figure 4.14: Gene set enrichment analysis of significant CpGs related to mammalian maximum lifespan in blood. The gene level enrichment was done using GREAT analysis using human background. The background probes were limited to the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). Human GWAS enrichment was calculated by a hypergeometric test of the top 5% genomic regions involved in GWAS of complex traits-associated genes with the top lifespan-related gene regions in our analysis. The biological processes were reduced to parent ontology terms using the "rrvgo" package. Input: Lifespan hypo/hyper, 500/500 CpGs; Lifespan (AdjWeight) hypo/hyper, 500/500. In each panel, the columns with no significant terms were removed to simplify the figure. Panels only show entries below a p-value threshold of $p<1x10^{-4}$.

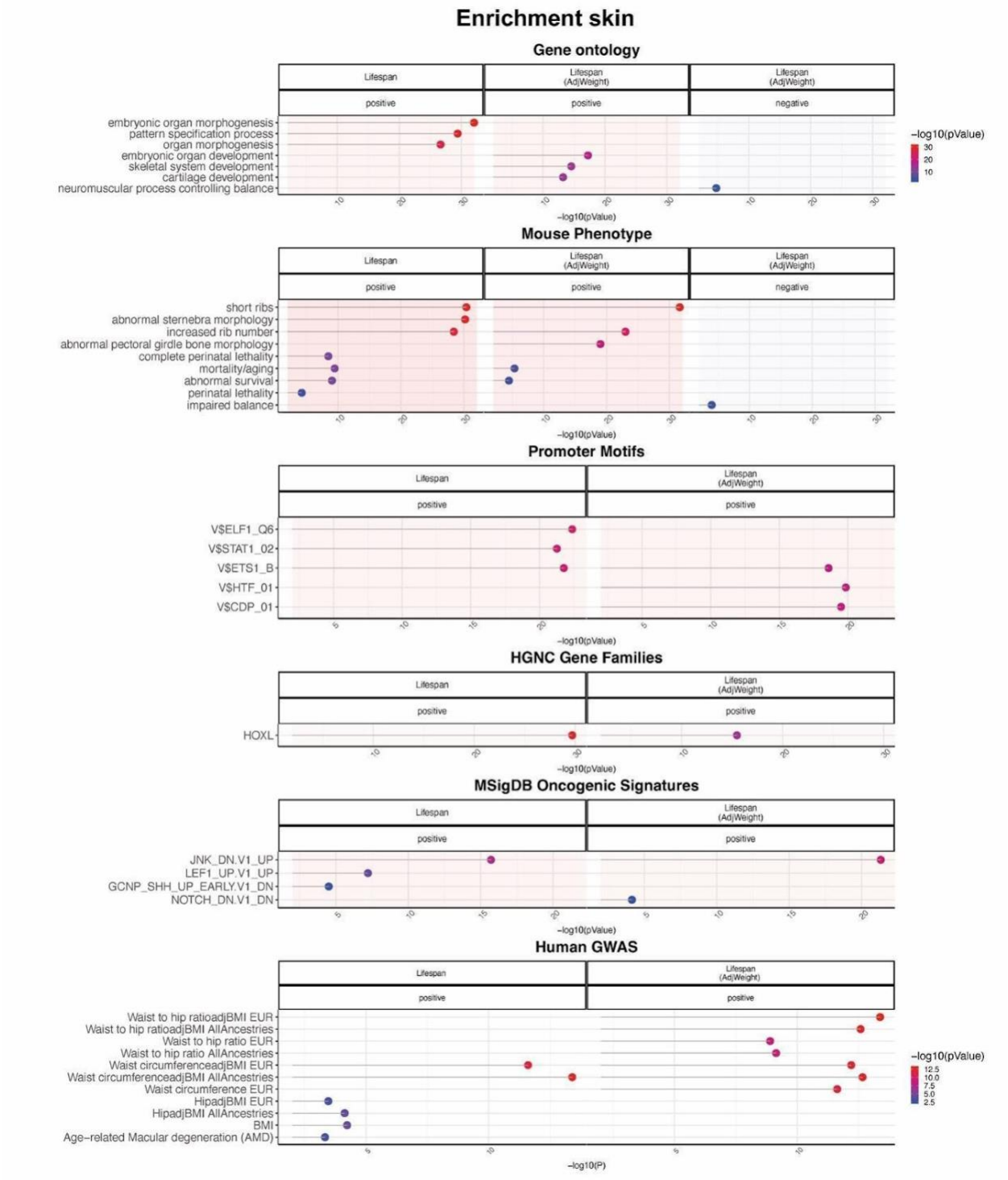

Figure 4.15: Gene set enrichment analysis of significant CpGs related to mammalian maximum lifespan in skin. The gene level enrichment was done using GREAT analysis using human background. The background probes were limited to the set of EWAS background CpG probes (28,318) consistent with the methods section (mappable to humans and mice and correlation with calibration exceeds 0.8). Human GWAS enrichment was calculated by a hypergeometric test of

the top 2.5% genomic regions involved in GWAS of complex traits-associated genes with the top lifespan-related gene regions in our analysis. The biological processes were reduced to parent ontology terms using the "rrvgo" package. Input: Lifespan hypo/hyper, 500/500 CpGs; Lifespan (AdjWeight) hypo/hyper, 500/500; Lifespan (AdjPhylo) hypo/hyper, 12/22; Lifespan (AdjPhyloWeight) hypo/hyper, 38/13. In each panel, the columns with no significant terms were removed to simplify the figure. Panels only show entries below a p-value threshold of $p<1x10^{-4}$.

1.1. Chapter acknowledgements

This chapter is a slightly modified version of my first-author paper published in the preprint, bioRxiv (C. Li et al., 2021), and has been reproduced here. Another version is being under review for a scientific journal.

CHAPTER 5

5. A novel L_0 regularized Poisson graphical model for RNA-seq data

This chapter presents a sparse Poisson graphical model. When using high-throughput sequencing technologies to measure gene expression, researchers are often interested in constructing a sparse network model. One established approach, Poisson Graphical LASSO (Allen & Liu, 2012), is implemented by fitting L_1-regularized regression models. However, it is well known that L_0-regularized regressions produce more parsimonious and accurate models, compared to L_1-regularized methods. However, direct L_0 norm regularization is difficult to estimate because of function convexity. L_2-norm penalization, on the other hand, is convex and stable, but in lack of sparsity. In this research we developed a new L_0 based Poisson graphical model, using cyclic coordinate-wise broken adaptive ridge (BAR) regression. This graphical model combines the benefits of both L_1 and L_2 penalization models and achieves an L_0-equivalent penalization. Performance of the model is evaluated and compared with some existing methods on both simulated and real data.

This chapter is organized as follows. In Section 5.1, we explain the motivation for a log-linear graphical model (LLGM) in general. In Section 5.2, we describe the proposed L_0-LLGM methodology in detail. In this section we also define notations necessary for graphical model constructions, and review steps of LLGM model. In the Subsection 5.2.3, we introduce a regularization parameter selection procedure for graphical models based on a stability algorithm. Then in Section 5.3, we compare performances of our proposed L_0-LLGM and standard model L_1-LLGM by simulating RNA-seq type of data from a few known network structures. Finally, in Section 5.4, we demonstrate a model application to a real world data set, kidney renal clear cell

carcinoma (KIRC) micro-RNA (miRNA) data from the Cancer Genome Atlas (TCGA) (Collins and Barker, 2007).

5.1. Motivations for a new Poisson graphical model

High dimensional analysis in gene expression studies often requires identifying associations between genes. Scientists are usually interested in a sparse network, which provides biologists with insights into possible pathways from particular groups of genes (Dobra et al., 2004; J. Friedman, Hastie, & Tibshirani, 2008; Meinshausen & Bühlmann, 2006). Multivariate Gaussian graphical models (J. Friedman et al., 2008; Meinshausen & Bühlmann, 2006). have been widely used to model continuous microarray data, since log ratios of the microarray gene expressions are approximately normally distributed after normalization. More recently, next generation high-throughput sequencing (RNA-seq) has become a popular data collection method for expression analysis (Dillies et al., 2013). Because RNA-seq gene expression data consist of counts of sequencing reads for each gene, researchers sought discrete probabilistic models, in favor of continuous Gaussian models, to describe the RNA-seq data (Allen & Liu, 2012; Chiquet, Robin, & Mariadassou, 2019; Choi et al., 2017; Gallopin, Rau, & Jaffrézic, 2013; Imbert et al., 2018; Srivastava & Chen, 2010; Witten, 2011). Some of these previous studies address zero-inflated Poisson distributions (Choi et al., 2017), whereas others focus on multivariate Poisson models (Chiquet et al., 2019).

Owing to restrictions in assumptions imposed by some joint models, some seek to build network models using neighborhood selection (Allen & Liu, 2012; Meinshausen & Bühlmann, 2006). A key advantage of neighborhood selection, in contrast to a joint distribution model, is that each neighborhood sparse estimation can be done simply by a multivariate log-linear regression,

and the regression model can be regularized conveniently by popular regularization methods such as L_1-regularized Lasso. Neighborhood network selections assume a pair-wise Markov property (Lauritzen, 1996): conditional on all other variables, each variable follows a Poisson distribution, and is estimated locally through neighborhood selection (Meinshausen and Bühlmann, 2006) by fitting L_1-regularized log-linear models (Allen and Liu, 2012). This Poisson graphical model based on neighborhood selection by Allen and Liu (2012) was recognized as one of the recent studies of graphical modeling specifically for discrete data with Poisson distributions, and it addresses conditional variable relationships without the need for a joint discrete distribution (Gallopin et al., 2013; Choi et al., 2017; Imbert et al., 2018; Chiquet et al., 2019). For the rest of the chapter, we address this model as L_1 log-linear graphical model (L_1-LLGM).

However, the L_1-LLGM method has some pitfalls because L_1 regularization is known to lack oracle properties and tends to include unwanted noise variables (Zou, 2006; Zou and Zhang, 2009; Zhang, 2010). Consequently, the resulting estimated network is often not sparse enough when compared with the true underlying network structure. To mitigate this issue, Allen and Liu (2012) introduced a threshold to filter out small coefficients retained by L_1-regularized log-linear regressions. We demonstrate in simulations that the inferred network is not robust with respect to the threshold level, and can sometimes have a very poor performance when a suboptimal threshold level is used. Unfortunately, no practical guidance is available in the literature on how to choose an appropriate threshold level for a given data set. In subsequent parts of the chapter, we employ the same assumptions (local Markov property) and settings (normalization, power transformation) of the L_1-LLGM, and propose a more refined estimation method for this model by adopting an L_0-equivalent regularization.

We frame the goal of this chapter as an improvement over L_1-LLGM. To this aim, we developed and implemented an approximate L_0-regularized log-linear graphical model (L_0-LLGM) for constructing sparse gene network from RNA-seq count data. We consider L_0 regularization because it generally yields higher true-positive estimations than L_1 regularization and has been shown to be more accurate for feature selection and parameter estimation (Lin et al., 2010, 2020; Shen et al., 2012, 2013). Because exact L_0 regularization is computationally non-deterministic polynomial-time hardness (NP-hard) and only feasible for low dimension data, we adapt the recently developed broken adaptive ridge (BAR) method to approximate $\ell 0$regularization. Defined as the limit of an iteratively reweighted L_2-regularization algorithm, the BAR method is an approximate L_0-regularization method that enjoys the best of L_0 and L_2 regularizations with desirable selection, estimation, and grouping properties (Dai et al., 2018, 2020; Zhao et al., 2018, 2020; Kawaguchi et al., 2020b).

These desirable properties are important to our network analysis, as they offer a theoretical advantage of L_0 regularization in our Poisson graphical model. Similar to L_1-LLGM, our proposed L_0-LLGM assumes a pair-wise Markov property and estimates gene network structures through a local sparse LLGM that evaluates conditional network correlations to each node. Specifically, at each step, a regularized Poisson log-linear model is fitted on one node, using BAR regularization to introduce sparsity. Nonzero coefficients estimate edges extending from the node. The stability approach to regularization selection (StARS) method (Liu et al., 2010) is used to select the regularization tuning parameters for the graphical model. Our empirical studies suggest that the proposed L_0-LLGM generally produces network structures closer to the true structure than those of L_1-LLGM, as measured by receiving operating characteristic (ROC) curves.

5.2. Methodology of graphical models in the context of gene expression data

5.2.1. Data and notations

We define matrix X as the design matrix, where columns are variables and rows are samples. Matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_p)$ is $n \times p$, where $\boldsymbol{X}_j (j = 1, \dots p)$ is the j-th column. Based on this design matrix, we aim to construct an undirected network model that would reveal conditional dependence between variables. It applies to count data that are assumed to have Poisson distributions. We then define the structure of the network as $\boldsymbol{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$. $\boldsymbol{V}$ is the set of all vertices in the network, where each vertex represents a variable (e.g., miRNA), quantified by a vector of the corresponding counts of aligned sequencing reads from each sample. $\boldsymbol{E}$ represents the set of all edges connecting certain vertices.

5.2.2. L_0 Regularized Log-Linear Poisson Graphical Model

We consider a log-linear Poisson graphical model, which characterizes conditional Poisson relationships by assuming pair-wise Markov properties (Lauritzen, 1996). Specifically, we assume that for each $j = 1, \dots, p$, the conditional distribution of column j, $\boldsymbol{X}_j | \boldsymbol{x}_k \; \forall k \neq j$, is

(Equation 5.1)

$$p\left(\boldsymbol{X}_j \middle| \boldsymbol{X}_k \; \forall k \neq j, \boldsymbol{B}\right) \sim Poisson(e^{\Sigma \beta_{jk} x_k})$$

where the intercept term bj0 is not included in the model, as we assume at this point RNA-seq data have been adjusted for sequencing depth in normalization steps, and $\boldsymbol{B} = (\beta_{jk}, \forall k \neq j \in V)$ is a $p \times p$ adjacency matrix with each row vector of off-diagonal elements storing the corresponding log-linear Poisson regression coefficients.

The first step to neighborhood network selection method is to infer graphical networks by fitting the mentioned log-linear Poisson regression for every node $j, j = 1, \dots, p$. Specifically, at each

neighborhood selection step j ($j = 1, \ldots, p$), we determine only the potential edges connecting node j to all other nodes in the network. In addition, we couple the neighborhood selection method with the BAR, an approximate L_0-regularization method, to induce sparsity as detailed in the following algorithm.

For each $j, j = 1, \ldots, p$, we begin with an initial L_2-regularized (ridge) estimator of $\beta_{\neq j,j}$,

(Equation 5.2)

$$\hat{\beta}^{(0)}_{\neq j,j} = \underset{\beta_{\neq j,j}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[X_{ij}\left(\boldsymbol{X}_{i,\neq j}\beta_{\neq j,j}\right) - \exp\left(\boldsymbol{X}_{i,\neq j}\beta_{\neq j,j}\right) + \zeta_n \sum_{k \neq j} \beta^2_{jk} \right] \right\}$$

where the first term is the -2log likelihood for the j-th log-linear Poisson regression model, $\beta_{\neq j,j}$ is a vector of $p-1$ corresponding regression coefficients, and ζ_n is the ridge-regularization parameter. This initial step tuning parameter serves the purpose of giving iterative step a warm start. The BAR estimator defined hereunder has been shown to be robust for different choice of ζ_n in various model settings [see, e.g., Kawaguchi et al. (2020)—Figure 1 and N. Li, Peng, Kawaguchi, Suchard, and Li (2021)—Figure 7]. For a reasonable initial step estimation, we have set ζ_n to log (n), where n is the sample size. We then subsequently update the estimator of $\beta_{\neq j,j}$ by fitting reweighted L_2-regularized regressions with a tuning parameter λ_n:

(Equation 5.3)

$$\hat{\beta}^{(s)}_{\neq j,j} = \underset{\beta_{\neq j,j}}{\operatorname{argmin}} \left\{ -2l(\boldsymbol{\beta}) + \lambda_n \sum_{i=1}^{n} \frac{\beta^2_{jk}}{\left|\hat{\beta}^{(s-1)}_{jk}\right|^2} \right\}, s = 1,2, \ldots$$

The BAR (Kawaguchi et al., 2020a) estimator of b61⁄4j, j is defined as,

(Equation 5.4)

$$\hat{\beta}_{\neq j,j} = \lim_{s \to \infty} \hat{\beta}^{(s)}_{\neq j,j} .$$

The BAR estimator has been shown to possess the oracle properties in the sense that with large probability, it estimates the zero coefficients as 0's and estimates the non-zero coefficients as well as the scenario when the true sub-model is known in advance and a grouping property that highly correlated variables are naturally grouped together with similar coefficients (Kawaguchi et al., 2021).

A nonzero element in coefficient estimate vector $\hat{\beta}_{\neq j,j}$ indicates that there is an estimated network connection (edge) between the corresponding node j and one of the nodes $1, \dots, p \neq j$. The estimators $\hat{\beta}_{\neq j,j}, j = 1, \dots, p,$, provide estimates of the off-diagonal elements of adjacency matrix $\boldsymbol{B}$. Diagonal elements of $\boldsymbol{B}$ can be set to either missing or unity, since it is not meaningful to evaluate a node's relationship with itself. Note that $\boldsymbol{B}$ is also not necessarily symmetric, as fitting regressions on element *i* and *j* does not guarantee the same zero or nonzero coefficient corresponding to the same node. To deal with this non-symmetric issue, we chose to estimate based on the union of network edge constructions,

(Equation 5.5)

$$\hat{A}_{jk}(p) = max\{|sign(\widehat{\boldsymbol{B}}(p)_{jk})|, |sign(\widehat{\boldsymbol{B}}(p)_{kj})|\} \; \forall j \neq k.$$

Theoretically, whether to use the union or intersection of each network edge based on its two neighborhood selections concerning its two nodes is asymptotically identical (Meinshausen & Bühlmann, 2006). This less conservative approach of estimating by unions remains consistent with previous neighborhood selection Poisson graphical model literature (Allen & Liu, 2012). In other words, if either one of the two local log-linear regressions concerning the two nodes *i* and *j* produces a nonzero estimate, it implies conditional dependency. Consequently, the network estimate specifies an edge between nodes i and j. Therefore, estimated adjacency matrix is always

symmetrical. Estimated coefficients, $\widehat{\boldsymbol{B}}$, is then transformed to an adjacency matrix, $\hat{A}$, by simply changing all nonzero estimates to 1, namely, $\hat{A} = sign|\widehat{\boldsymbol{B}}|$.

Lastly, we note that for each $j, j = 1, \ldots, p$, the BAR estimator $\hat{\beta}_{\neq j,j}$ is defined as the limit of a sequence of reweighted ridge estimators. In a numerical implementation, one will stop the BAR iterations for $\hat{\beta}_{\neq j,j}$ when a prespecified convergence criterion is met. In our implementation, the algorithm stops at step s when $\max_{k \neq j} \left| \hat{\beta}_{jk}^{(s)} - \hat{\beta}_{jk}^{(s-1)} \right| < a$, and we set $\hat{\beta}_{\neq j,j} = \left\{ \hat{\beta}_{jk}^{(s)} I\left(\left| \hat{\beta}_{jk}^{(s)} \right| > a \right), k \neq j \right\}$, where a is the convergence criterion threshold, which can be set to a reasonably small value, such as 1×10^{-18}. Our empirical studies indicate that one may use a slightly larger value to reduce the number of iterations with essentially no difference in the resulting estimator. We set $a = 1 \times 10^{-16}$ as the default value in our R implementation. We emphasize that threshold a is not a regularization parameter. It is comparable with the stopping rule for a Lasso gradient descent implementation. It is purely for implementing the computer algorithm, as it serves as a stop mechanism for numerical convergence. This is not to be confused with the artificial threshold in L1-LLGM (Allen & Liu, 2012; Wan et al., 2016), which, in the R package, was imposed after Lasso gradient descent stopping rule, effectively "weeding out" small, but converged, Lasso coefficients.

5.2.3. Selecting regularization parameters through StARS criterion

The sparsity and performance of the network largely depend on the regularization parameter λ_n in (Equation 5.3, which directly determines the number of estimated edges that stay in the network. Note that most of the popular data-driven tuning parameter selection methods such as Akaike's information criteria (AIC), Bayesian information criteria (BIC), and cross-validation require

finding the log likelihood of the joint distribution, which all local neighborhood log-linear Poisson models do not have. Thus, we opt to select regularization parameters utilizing StARS (Liu, Roeder, & Wasserman, 2010). The StARS selection criterion selects the regularization parameter based on given model stability. It does so by subsampling rows, without replacement, into blocks of equal sizes.

Specifically, let K be the number of subsamples that we draw, and let X^k be a subsample from the design matrix X, where $k \in \{1, \dots, K\}$. Liu et al. (2010) suggest that, in order for assumptions of StARS algorithm to be met, a reasonable choice of the subsample size is $b = \lfloor 10\sqrt{n} \rfloor$. In our case of neighborhood Poisson graphical model, individual full models are fitted on each subsample. For any edge between two given vertices, we will have obtained K estimates on the same edge, each from a subsample already mentioned. First, we define an inverse of the tuning parameter $\Lambda = 1/\lambda_n$. For any edge connecting vertices s and t, let the estimate from subsample S_j, using regularization parameter Λ, be $\psi_{\mathrm{st}}^{\Lambda}(S_j)$. $\psi_{\mathrm{st}}^{\Lambda}(S_j) = 1$ if there is an edge between s and t, and $\psi_{\mathrm{st}}^{\Lambda}(S_j) = 0$ if the model does not estimate that there is an edge at (s, t). The stability of model predictions on this specific position is then given by

(Equation 5.6)

$$\hat{\theta}_{st}(\Lambda) = \frac{1}{N} \sum_{j=1}^{N} \psi_{\mathrm{st}}^{\Lambda}(S_j).$$

A potential issue here with estimator $\hat{\theta}_{st}(\Lambda)$ is that the measure is not monotonic, rendering future model assessment and comparisons difficult. The model is stable when estimates from different subsamples all tend to give a value of 1 or 0. In other words, $\hat{\theta}_{st}(\Lambda)$ is the most stable when it is close to 0 or 1, and the least stable when it is close to 0.5. Therefore, we use a monotonized stability measure,

(Equation 5.7)

$$\hat{\xi}_{st}(\Lambda) = 2\hat{\theta}_{st}(\Lambda)(1 - \hat{\theta}_{st}(\Lambda)).$$

Lastly, an overall stability measure is then calculated by evaluating the mean of all edge-specific instabilities,

(Equation 5.8)

$$\widehat{D}_{st}(\Lambda) = \frac{\sum_{s<t} \hat{\xi}_{st}(\Lambda)}{\binom{p}{2}}.$$

When Λ is close to 0, meaning regularization parameter λ is large, L_0-LLGM produces an empty graph. As all subsample estimates are sparse, the instability shall approach 0. As Λ increases, the subsample networks become denser and more volatile. Instability consequently increases till it peaks. $\widehat{D}_{st}(\Lambda)$ will start decreasing as the regularization parameter becomes smaller and the networks become dense. As the networks become almost fully connected, the instability measure will again approach 0, since all subsamples give similar estimates. Therefore, instabilities are expected to have a bell shape when plotted against penalization parameter. Authors who proposed StARS criterion also suggested a way to select the optimal sparsity given the least instability. Users first need specify an instability threshold, γ. Then the algorithm should select the largest penalization parameter, that is, the sparsest network, with instability score below or equal to γ.

The performance instability criterion is supported theoretically by the Theorem of Partial Sparsistency (Liu et al., 2010), which states that under suitable regularity conditions, the estimated set of edges is expected to contain the set of edges in the true underlying model as n approaches infinity.

We developed an R package for implementing L_0-LLGM, which can be found at repository https://github.com/caeseriousli/prBARgraph.git.

5.3. Simulations

In this section, we demonstrate the performance of the BAR Poisson graphical model (L_0-LLGM) versus L_1 Poisson graphical model (L_1-LLGM) through simulations. To evaluate model fit, we measure prediction accuracy by the true-positive rates and false-positive rates. The true-positive rate is defined as the portion of correctly predicted edge out of total number of edges predicted. For instance, if a predicted network has a total of 80 edges, out of which 40 exist in the underlying true network, then the true-positive rate is 40/80 = 0.5 in this case. False-positive rate, however, is calculated by dividing the number of incorrectly predicted edges by the total number of non-existing edges in the true network.

5.3.1. Simulating correlated Poisson networks

To generate simulation data for model comparison, we adapt the same method introduced in Allen and Liu (2012). Again, let n be the number of observations and p the number of elements (genes). We first generate independent Poisson samples: Y, a $n \times (p + p(p-1)/2)$ matrix, where$Y_{ij} \sim Poisson(\lambda_{true})$. Then we randomly generate a noise term E, an $n \times p$ matrix where $E_{ij} \sim Poisson(\lambda_{true})$.

Furthermore, using the underlying true network, we construct a structure matrix,

(Equation 5.9)

$$\widehat{D}_{st}(\Lambda) = \frac{\sum_{s<t} \hat{\xi}_{st}(\Lambda)}{\binom{p}{2}}.$$

where A is the adjacency matrix corresponding to the network, $tri(A)$ is a vectorized, $\left(p \times \frac{p-1}{2}\right) \times 1$, upper triangular part of adjacency matrix A, and $\mathbf{1}_{(p)}$ here is a column vector of

which each element is equal to 1. The purpose of the identity vector is to expand $tri(A)$ into a $p \times \frac{p-1}{2}$ matrix, which is used to calculate element-wise product with a permutation matrix, $\boldsymbol{P}$, with dimensions $(p(p-1)) \times p$. The permutation matrix is constructed by permuting indices of all possible pairs of vertices across its rows. For example, if the first row of the permutation matrix, $\boldsymbol{P}$, represents an edge connecting node number 1 and node number 2, then the first two elements of the first row, which contains a total of p elements, will be 1. The rest of elements in the first row are 0. Concordantly, P has $p(p-1)/2$ rows because a network can potentially have a total of distinct $p(p-1)/2$ edges. Note that the order of permutations in P have to match the order we expand the adjacency matrix, namely, $tri(A)$. In addition, denotes the block matrix structure with the $p \times p$ identity matrix on the left. Finally, we simulate the design matrix by $\boldsymbol{X} = \boldsymbol{YB} + \boldsymbol{E}$.

5.3.2. Model comparison

When compared with nondiscrete models, such as graphical Lasso (J. Friedman et al., 2008), the L_1-LLGM model has already been numerically demonstrated to have as good or better prediction accuracy for simulated Poisson data (Allen & Liu, 2012). In this chapter, as the major innovation is an L_0 BAR regularization, we will focus on comparing L_0-LLGM with L_1-LLGM. We will move on to adopt simulation setup similar to Allen and Liu (2012). Specifically, we simulated RNA sequencing data based on two common network topologies, hub and scale free. The data are randomly generated using methods introduced in Section 5.3.1. For each topology, we have constructed a network consisting of 50 nodes. For each topology we generate two data sets, with 200 and 500 independent samples, respectively. We further note that, sample sizes no greater than 500 in simulations are common for most RNA-seq studies (Chiquet et al., 2019; Choi et al., 2017; Gallopin et al., 2013; Imbert et al., 2018).

For each model, both the L_0-LLGM and L_1-LLGM methods are performed on the simulated data, using StARS criterion to determine the regularization parameters. For L_1-LLGM, we considered a set of four different values for the additional sparsity threshold ("th"), specified by L_1-LLGM implementations as a necessary step (Wan et al., 2016), to investigate its effects on the resulting estimated network. With both true-positive and false-positive measurements defined in the beginning of Section 3, we construct ROC curves to compare the performance of L_0-LLGM in comparison with L_1-LLGM. Figure 5.1 shows the ROC curves generated under two different topologies, scale free (Figure 5.1A) and hub (Figure 5.1D), each consisting of 50 nodes. We observe that L_0-LLGM consistently outperformed L_1-LLGM, especially in high specificity regions. The advantage of L_0-LLGM is more evident for hub topology. Furthermore, it is clear that the performance of L_1-LLGM can vary greatly depending on the choice of its sparsity threshold. The optimal choice of this threshold depends on the underlying topology and sample sizes. For any given false-positive rate, L_0-LLGM yields a model with more correctly estimated connections. L_1-LLGM, in contrast, could potentially lose nodes that are important to the structure of the network shown in Figure 5.1.

Lastly, we validate the mentioned findings in replications. Owing to limitations of ROC plot visualizing multiple network fits, we summarize 40 replications in box plots. In Figure 5.2 and Figure 5.3, L_0-LLGM and L_1-LLGM are fitted to 40 randomly generated data sets from scale-free and hub topologies, respectively. At each replication, both the topology and data set are randomly generated, and each data set has sample size n=500.

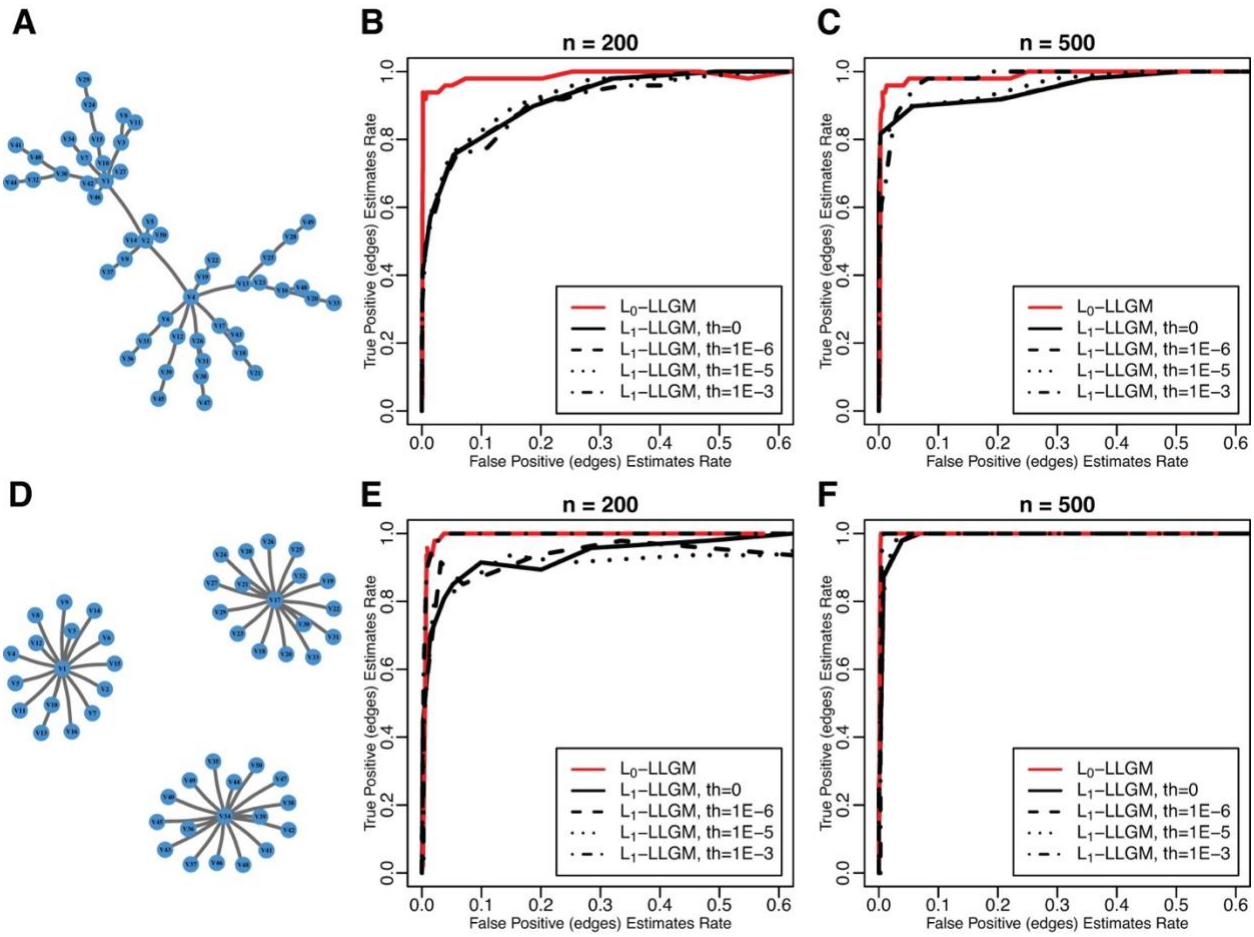

Figure 5.1: Simulation study for two network topologies: (A) scale free and (B) hub. For each network structure, we generated two data sets with two different number of observations, 200 and 500. A sequence of $\ell 0$ or $\ell 1$ penalization parameters was used to fit the modes on each data set. Predictions were evaluated by calculating true-positive and false-positive rates. These rates from both models were plotted for model comparisons (B, C, E, F). (B) and (C) Are two data sets, based on a scale-free network in (A), with simulated sample sizes equal to 200 and 500, respectively, while (E) and (F) are the same sample sizes based on a hub network in (D).

5.4. Application of L_0-LLGM to KIRC MIRNA-seq data

High throughput sequencing (second generation RNA sequencing) returns millions of short reads of RNA fragments, which have varying lengths ranging from ~25 to possibly 300bp paired-end reads (Chhangawala, Rudy, Mason, & Rosenfeld, 2015). These reads are usually mapped to the genome and the data are in the form of non-negative counts of the RNA fragment reads (Witten, 2011). LLGM models can be applied to any data that are assumed to have Poisson distributions.

For RNA-seq data specifically, a normalization pipeline is required before data analysis. For comparison purposes, we follow the same normalization pipeline as Wan et al. (2016); Allen and Liu (2012), which consists of the following major steps: (1) adjusting for sequencing depth, (2) biological entities (e.g., genes, miRNAs) with low counts or low variances are filtered out, (3) vectors with potential over-dispersion are transformed using a power transformation to transform the data closer to Poisson distribution (J. Li, Witten, Johnstone, & Tibshirani, 2012; Wan et al., 2016). The normalization steps can be performed by R package XMRF (Wan et al., 2016). We defer the detailed procedures and justifications of this specific pipeline for RNA-seq Poisson graphical models to Wan et al. (2016).

We then fitted the proposed method on the KIRC miRNA data set from The Cancer Genome Atlas. The data set was downloaded from TCGA data portal (https://portal.gdc.cancer.gov) (Collins & Barker, 2007). It contains 1881 miRNAs and 616 samples. Before the normalization pipeline, we filtered out miRNAs that have all zero read counts throughout all samples, resulting in 1502 miRNAs left (20.15% of miRNAs with low counts). Then we normalize the rest of the data using XMRF package developed for L_1-LLGM (Wan et al., 2016). For demonstration purposes of this chapter, we specify the R package to keep top 100 miRNAs with the most variance (i.e., look at top miRNAs that vary the most). Minimum read count is set to be no less than 20, the suggested default (Allen & Liu, 2012; Wan et al., 2016). This keeps ~6.7% of miRNAs. We then move on to focus on conditional relationships between these 100 miRNAs with the largest variance and reasonable read counts. We then fit L_0-LLGM using our R package, along with L_1-LLGM, implemented by XMRF (Wan et al., 2016), with a StARS instability threshold of $\gamma = 0.01$ (choosing the largest regularization while maintaining at least "99% stability"). Figure 5.4 and Figure 5.5 show the resulting L_0-LLGM and L_1-LLGM network estimates, respectively. Figure 5.5

contains four panels, each with a different L_1-LLGM artificial threshold. Network estimates could be drastically different depending on the threshold. Table 5.1 provides miRNA annotations for use of node numbers in Figure 5.4 and Figure 5.5.

It is observed from Figure 5.4 and Figure 5.5 that the two model results reveal some similar structures, including the hub surrounding center, mir-10b (node 25). However, L_0-LLGM produces a less visually "chaotic" network, in comparison with L_1-LLGM. For instance, L_0-LLGM outlines a clean scale-free topology with minimal cyclic loops. L_1-LLGM, however, frequently exhibits loops even in sparse network estimates, possibly due to unwanted noises from L_1 regularization. From Figure 5.5, as we increase the "artificial threshold" for L_1-LLGM used in XMRF package, to some extent it helps reducing these noise edges.

However, during this process, we observe that this user-imposed threshold also filtered out lower degree nodes (weaker signal), such as the hub miRNAs surrounding node 25. For example, in Figure 5.5A1, with no artificial threshold, L_1-LLGM identifies hub center node 25 (mir-10b), which agrees with L_0-LLGM in Figure 5.4. As the threshold increases, plots in Figure 5.5A2–B2 show a decreasing degree in hub center node 25. In Figure 5B2, almost entire hub is filtered out by this threshold along with noise. This observation parallels to the simulation section (Figure 5.1), where the true-positive rates can be reduced by the artificial threshold, losing important network structures. These preliminary observations suggest that L_0-LLGM is potentially more capable of separating signal from noise, which is consistent with our simulation results depicted in Figure 5.1.

Although a graphical model alone is not enough to make any further conclusions on gene interactions inference, we focus, in particular, on highly connected miRNAs (i.e., hub nodes). Some of the hub miRNAs revealed by the L_0-LLGM network were previously known to be associated with each other, and with certain cancers. For example, the center of the largest hub,

gene mir-10b in Figure 5.4 is known to be associated with cancers such as bladder cancer and proteoglycans cancer. Based on literature studies, this RNA was known to be highly expressed in metastatic hepatocellular carcinomas, in contrast to those without metastasis (Ma et al., 2010). Our network results are based on the data from patients with adenomas and adenocarcinomas from project KIRC. It is connected to numerous miRNAs, including several cluster centers known to be associated with cancer suppressing. RNA named hsa-let-7b, for example, identified as a sub-cluster connected to the hub center mir-10b, a previously known putative cancer suppressor, is found to play a key role in chemoresistance in renal cells from carcinoma cases (Peng, Mo, Ma, & Fan, 2015). Together with another cluster center RNA, named miR-126 and hsa-let-7b are both identified as crucial biomarkers for identifying renal cell carcinoma (Carlsson et al., 2019; Jusufović et al., 2012; Yin et al., 2014). Our graphical model successfully identifies important miRNAs that align with published biological findings regarding such miRNAs.

We also performed additional analyses using different StARS instability thresholds $\gamma = 0.005$ and $\gamma = 0.05$. The findings are consistent with what have been discussed previously for $\gamma = 0.01$ and thus not included here.

5.5. Discussion

We have proposed and implemented an approximate L_0-LLGM for constructing sparse gene network from RNA-seq count data. This approach uses a neighborhood Poisson graphical model, which offers a more comprehensive set of predictions, has less constraints on the Poisson distributions of each element, and is less sensitive to changes of individual genes, than a joint distribution model. Sparsity is achieved through the BAR penalization, a surrogate L_0 regularization with established oracle properties for selection and estimation. Our simulations in

Section 3 show that, in general, L_0-LLGM offers theoretically more accurate estimates than L_1-LLGM. It reaches a high level of true-positive rate faster, without accumulating a high rate of false estimates. L_0-LLGM also spares users the need of selecting an additional sparsity threshold after the regularization tuning parameter has already been selected by StARS. This brings more consistency and reproducibility to the graphical model.

Our simulations considered two types of network topologies, namely scale-free and hub topologies, and found that both L_0-LLGM and L_1-LLGM tend to perform better under scale-free topologies as compared with hub. However, because graphical models could potentially give drastically different results under various topologies, it would be of interest to consider more topologies in future studies.

5.6. Additional figures and tables

5.6.1. Figures

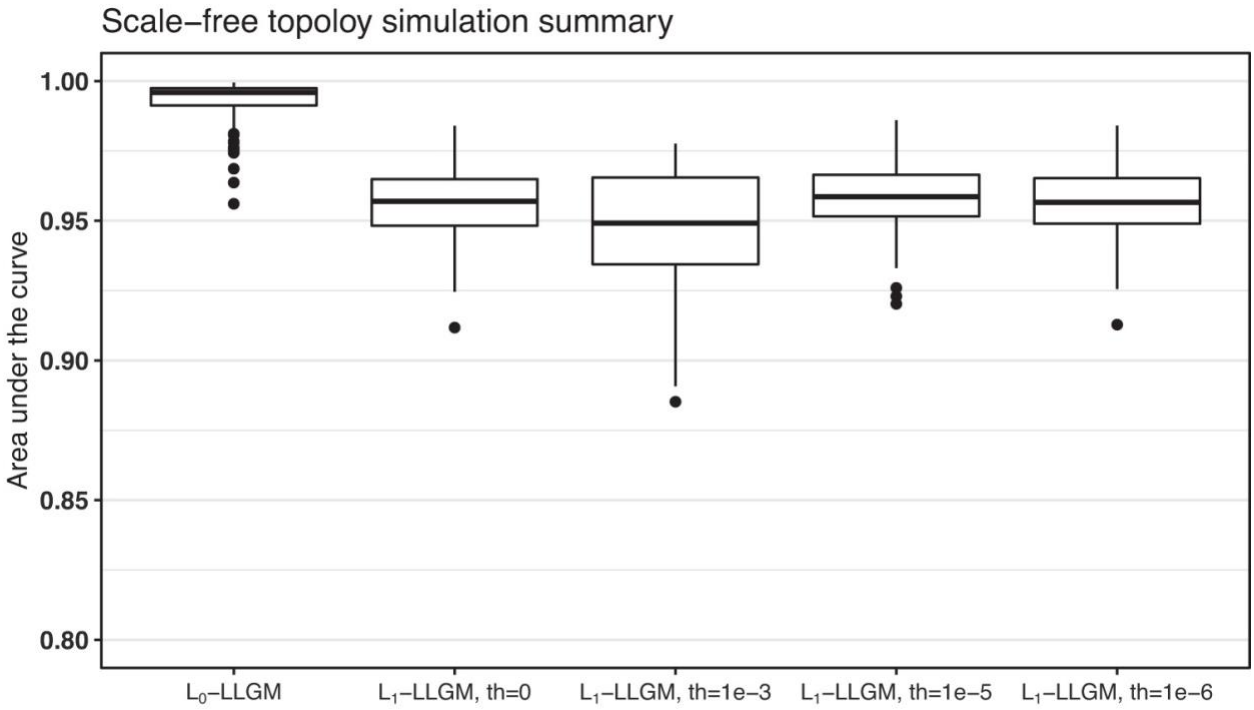

Figure 5.2: Simulation study for scale-free topology with sample size n=500. Topologies and data sets are randomly generated 100 times for each model. For all repetitions, area under the curve for true-positive rates and false-positive edge estimation percentages are summarized in box plots. Area under the curve is defined as the area under true-positive versus false-positive rate curve as regularization parameter increases, same as that of Figure 5.1. These repeated simulations are based on randomly generated scale-free topologies with 200 sample sizes and 50 number of nodes, corresponding to the same specifications in Figure 5.1B. Both the topology and data set are simulated randomly at each repetition

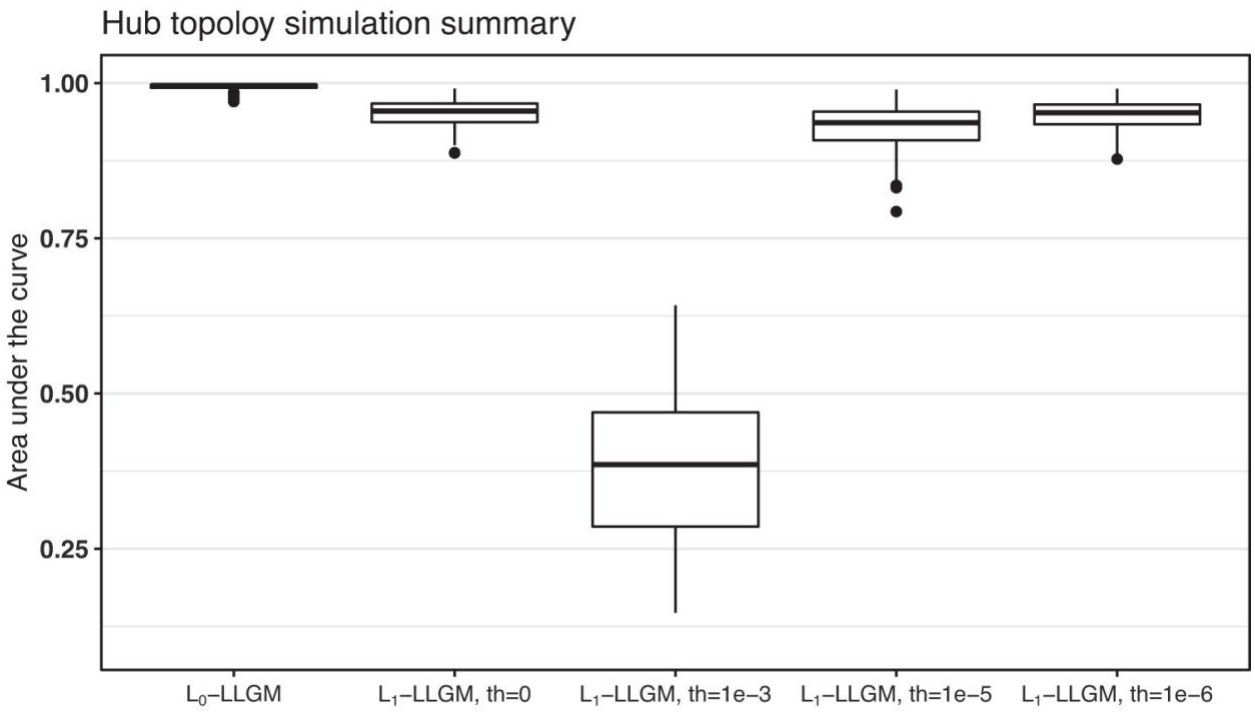

Figure 5.3: Simulation study for hub topology with sample size n=500. Topologies and data sets are randomly generated 100 times for each model. For all repetitions, area under the curve for true-positive rates and false-positive edge estimation percentages are summarized in box plots. Area under the curve is defined as the area under true-positive versus false-positive rate curve as regularization parameter increases, same as that of Figure 5.1. These repeated simulations are based on randomly generated hub topologies with 200 sample sizes and 50 number of nodes, corresponding to the same specifications as in Figure 5.1E. Both the topology and data set are simulated randomly at each repetition.

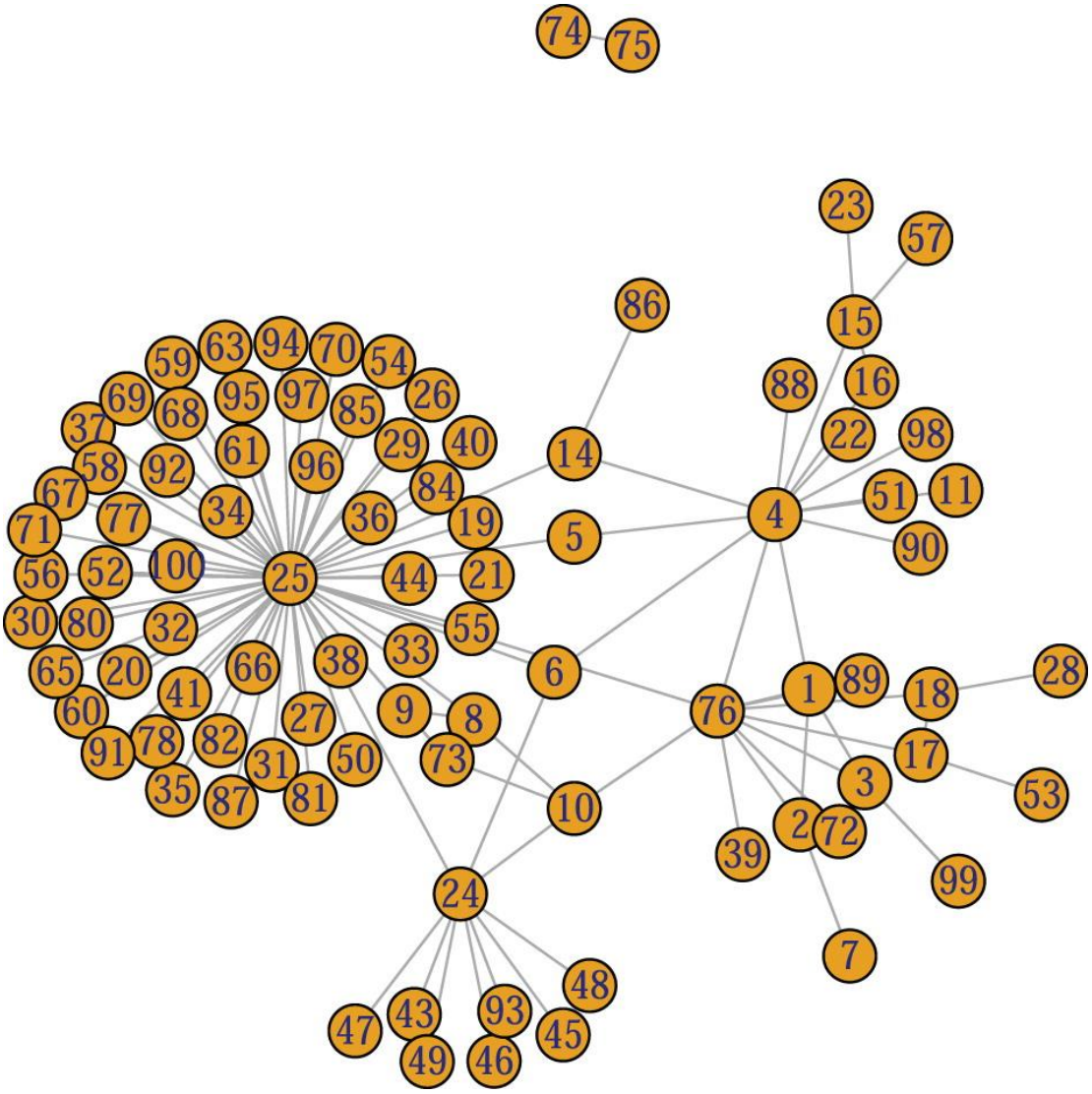

Figure 5.4: L_0-LLGM KIRC miRNA data: estimated network generated by fitting an L_0-LLGM model on KIRC miRNA data from TCGA database. The penalization parameter was chosen by setting a StARS estimation instability threshold of 0.01. KIRC, kidney renal clear cell carcinoma; L_0-LLGM, L_0-regularized log-linear graphical model; miRNA, micro-RNA; StARS, stability approach to regularization selection; TCGA, the Cancer Genome Atlas.

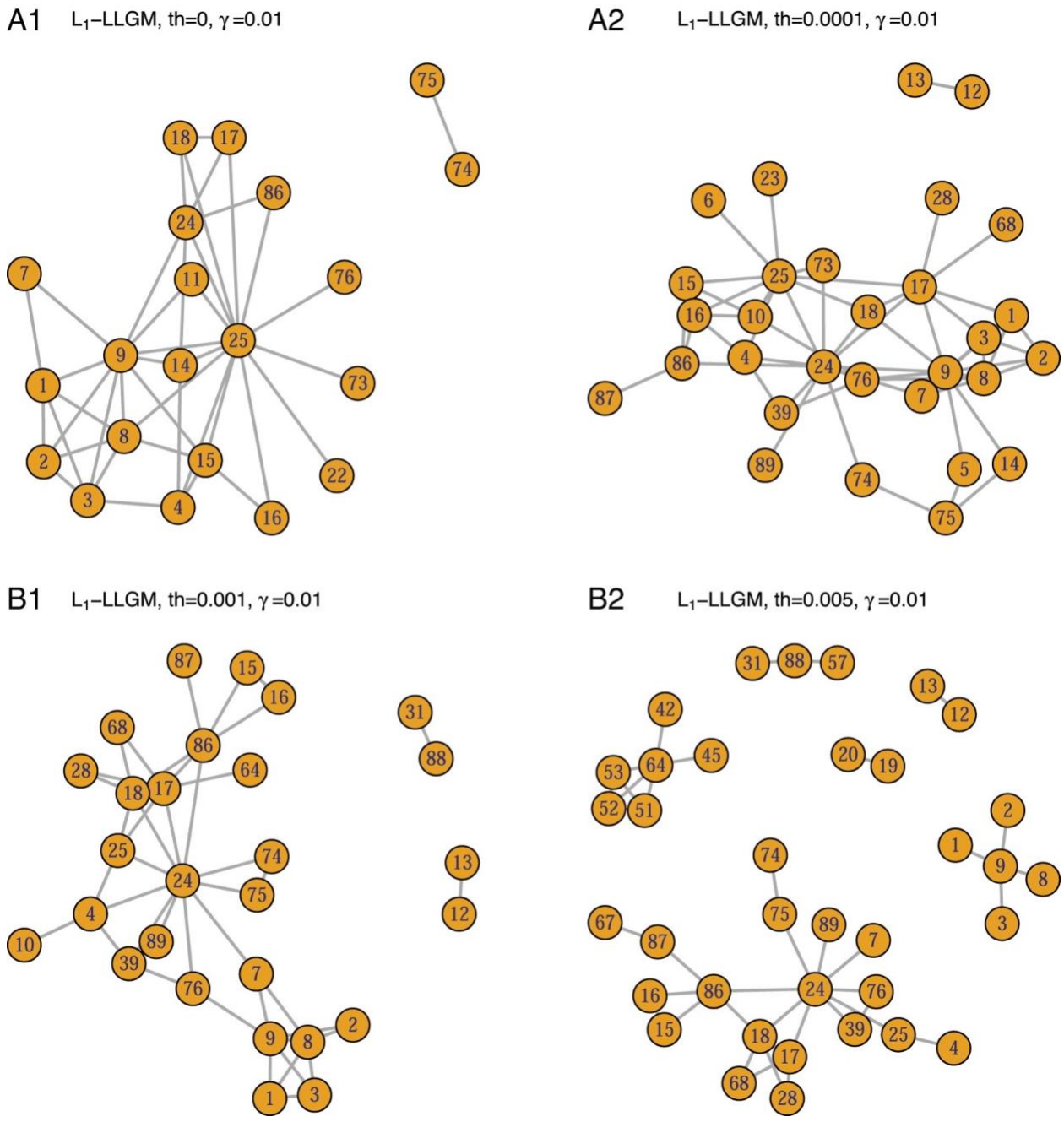

Figure 5.5: L_1-LLGM KIRC miRNA data: estimated network generated by fitting an L_1-LLGM model on KIRC miRNA data from TCGA database. The penalization parameter was chosen by setting an StARS instability threshold of 0.01. In addition, a further artificial threshold ("th") to fine tune the L_1-LLGM model. This figure shows four network estimates by varying the th threshold.

5.6.2. Tables

Table 5.1: Micro-RNA (miRNA) look-up table. Each ID in Figure 5.4 and Figure 5.5 correspond to an miRNA in this table.

ID	miRNA	ID	miRNA	ID	miRNA	ID	miRNA
1	hsa-let-7a-1	26	hsa-mir-1178	51	hsa-mir-124-1	76	hsa-mir-126
2	hsa-let-7a-2	27	hsa-mir-1179	52	hsa-mir-124-2	77	hsa-mir-1260a
3	hsa-let-7a-3	28	hsa-mir-1180	53	hsa-mir-124-3	78	hsa-mir-1260b
4	hsa-let-7b	29	hsa-mir-1181	54	hsa-mir-1243	79	hsa-mir-1262
5	hsa-let-7c	30	hsa-mir-1182	55	hsa-mir-1244-1	80	hsa-mir-1263
6	hsa-let-7d	31	hsa-mir-1185-1	56	hsa-mir-1244-2	81	hsa-mir-1264
7	hsa-let-7e	32	hsa-mir-1185-2	57	hsa-mir-1245a	82	hsa-mir-1265
8	hsa-let-7f-1	33	hsa-mir-1193	58	hsa-mir-1245b	83	hsa-mir-1266
9	hsa-let-7f-2	34	hsa-mir-1197	59	hsa-mir-1246	84	hsa-mir-1267
10	hsa-let-7g	35	hsa-mir-1199	60	hsa-mir-1247	85	hsa-mir-1268b
11	hsa-let-7i	36	hsa-mir-1200	61	hsa-mir-1248	86	hsa-mir-1269a
12	hsa-mir-1-1	37	hsa-mir-1203	62	hsa-mir-1249	87	hsa-mir-1269b
13	hsa-mir-1-2	38	hsa-mir-1204	63	hsa-mir-1250	88	hsa-mir-127
14	hsa-mir-100	39	hsa-mir-122	64	hsa-mir-1251	89	hsa-mir-1270
15	hsa-mir-101-1	40	hsa-mir-1224	65	hsa-mir-1252	90	hsa-mir-1271
16	hsa-mir-101-2	41	hsa-mir-1225	66	hsa-mir-1253	91	hsa-mir-1272
17	hsa-mir-103a-1	42	hsa-mir-1226	67	hsa-mir-1254-1	92	hsa-mir-1273c
18	hsa-mir-103a-2	43	hsa-mir-1227	68	hsa-mir-1254-2	93	hsa-mir-1273h
19	hsa-mir-105-1	44	hsa-mir-1228	69	hsa-mir-1255a	94	hsa-mir-1275
20	hsa-mir-105-2	45	hsa-mir-1229	70	hsa-mir-1256	95	hsa-mir-1276
21	hsa-mir-106a	46	hsa-mir-1231	71	hsa-mir-1257	96	hsa-mir-1277
22	hsa-mir-106b	47	hsa-mir-1234	72	hsa-mir-1258	97	hsa-mir-1278
23	hsa-mir-107	48	hsa-mir-1236	73	hsa-mir-125a	98	hsa-mir-128-1
24	hsa-mir-10a	49	hsa-mir-1237	74	hsa-mir-125b-1	99	hsa-mir-128-2
25	hsa-mir-10b	50	hsa-mir-1238	75	hsa-mir-125b-2	100	hsa-mir-1281

5.7. Chapter acknowledgements

This chapter is a slightly modified version of my first-author paper published in the Journal of Computational Biology (C. Z. Li et al., 2021), and has been reproduced here with the permission of the copyright holder, “Mary Ann Liebert, Inc. publishers does not require authors of the content being used to obtain a license for their personal reuse of full article, charts/graphs/tables or text excerpt.”

CHAPTER 6

6. Concluding remarks and future research considerations

6.1. Remarks for high dimensional inference for DNA methylation data

So far, we have solved the theoretical high dimensional prediction problem with elastic net, but it lacks some features that come with a standard linear model. It is currently not possible to construct confidence intervals or compute coefficient p-values for selected variables. The only inference step we can take is list the CpG sites selected by the models. It is, however, non-trivial to infer which variable is more or less significant, nor can we construct confidence intervals for the predicted values. While most machine learning engineers would stop at prediction, my future work as a statistician will involve some form of inference, to help us better understand the underlying mechanism of epigenetic aging. Recently, some statisticians have theorized a few models for high-dimensional inference, including desparsified lasso (Van de Geer, Bühlmann, Ritov, & Dezeure, 2014), and a selection-assisted partial regression (SPARES) (Fei, Zhu, Banerjee, & Li, 2019). Briefly, desparsified Lasso is based on Lasso model. It takes advantage of Karush-Kuhn-Tucker characterization of Lasso and compute an approximation to the inverse of $X^T X/n$. This combined with central limit theorem gives an estimated covariance matrix for estimated coefficients. The pros of this method include the fact that it does not require any random data splitting step, relatively easy computation load, and that it is related to Lasso penalization. One disadvantage, however, is that the method is non-trivial to be generalized to elastic net, inheriting all the flaws of Lasso discussed in Chapter 1. SPARES involves randomly separating data set into two sub-groups, one for variable selection and one for inference. In addition, as the algorithm makes inferences on selected variables, it also assigns p-values to unselected variables by including them to the design matrix one by one. Then it performs the data splitting iteratively to achieve stable p-value estimates.

Two major advantages of this method are that it provides unbiased estimates with normality assumption held, and it is compatible with many different variable selection methods. For example, one can use elastic net for variable selection step in the half sub-sampled data, and then draw inference using standard linear model on the other half. Consequently, a drawback of this algorithm is that it can be computationally expensive for large data sets. In future research, one may test both of the named methods on the mammalian data and assess the practicality of both models. The ability to make inference on epigenetic predictors will be crucial to future aging interventions and possible clinical trials.

6.2. Future research for L_0-regularized Poisson graphical model

It is worth noting that although the method described in Chapter 5 focuses on the $p < n$ case, the proposed methodology can be easily extended to high dimensional settings where $p > n$ by coupling the BAR penalization with a sure screening procedure (Barut, Fan, & Verhasselt, 2016; Fan & Lv, 2008; Xu & Chen, 2014; S. D. Zhao & Li, 2012). Combining the BAR penalization with a sure screening procedure for high dimensional settings and its statistical guarantees have been studied for a variety of models including linear model (Dai et al., 2018), generalized linear models (N. Li et al., 2021), and survival models (Kawaguchi et al., 2020; H. Zhao et al., 2018; H. Zhao, Wu, Li, & Sun, 2019). Future studies are warranted to further investigate the empirical performance of the two-step procedure for network inference in high dimensional settings.

We acknowledge that, for model applications in RNA-seq data, the network in itself often is not enough to draw definitive inference on complex gene interactions. Often a network serves as a first step in identifying clusters, under the assumptions that genes interact with each other in hubs (N. Friedman, 2004). One can modularize clustered genes through methods such as dynamic tree cutting. These modules can subsequently be used for gene enrichment analyses (Langfelder

& Horvath, 2008). These subsequent procedures would all benefit from a proper graphical model such as the L_0-LLGM. In addition, as the BAR algorithm optimization progresses, it will become more feasible to implement an L_0-regularized regression, which possesses theoretical oracle properties, for DNA methylation data, resulting in superior variable selection and regression estimates.

7. Bibliography

Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N., & Warren, S. T. (2012). Age-associated DNA methylation in pediatric populations. *Genome research, 22*(4), 623-632.

Allen, G. I., & Liu, Z. (2012). *A log-linear graphical model for inferring genetic networks from high-throughput sequencing data.* Paper presented at the 2012 IEEE International Conference on Bioinformatics and Biomedicine.

Arneson, A., Haghani, A., Thompson, M. J., Pellegrini, M., Kwon, S. B., Vu, H., . . . Horvath, S. (2021). A mammalian methylation array for profiling methylation levels at conserved sequences. *Biorxiv*, 2021.2001.2007.425637. doi:10.1101/2021.01.07.425637

Arneson, A., Haghani, A., Thompson, M. J., Pellegrini, M., Kwon, S. B., Vu, H., . . . Lu, A. T. (2022). A mammalian methylation array for profiling methylation levels at conserved sequences. *Nature Communications, 13*(1), 1-13.

Austad, S. N. (2010). Methusaleh's Zoo: how nature provides us with clues for extending human health span. *Journal of comparative pathology, 142 Suppl 1*(Suppl 1), S10-S21. doi:10.1016/j.jcpa.2009.10.024

Barut, E., Fan, J., & Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association, 111*(515), 1266-1277.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., . . . Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology, 12*(1), 1-13.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*(57), 289-300.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., . . . Gunderson, K. L. (2011). High density DNA methylation array with single CpG site resolution. *Genomics, 98*(4), 288-295.

Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S., & Vilain, E. (2011). Epigenetic predictor of age. *PloS one, 6*(6), e14821.

Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E., . . . Ophoff, R. A. (2009). The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PloS one, 4*(8), e6767.

Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., . . . Baccarelli, A. (2009). Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mechanisms of ageing and development, 130*(4), 234-239.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8*, 3-62.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, 13-60.

Booth, L. N., & Brunet, A. (2016). The Aging Epigenome. *Mol Cell, 62*(5), 728-744. doi:10.1016/j.molcel.2016.05.013

Carlsson, J., Christiansen, J., Davidsson, S., Giunchi, F., Fiorentino, M., & Sundqvist, P. (2019). The potential role of miR-126, miR-21 and miR-10b as prognostic biomarkers in renal cell carcinoma. *Oncology Letters, 17*(5), 4566-4574.

Chhangawala, S., Rudy, G., Mason, C. E., & Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology, 16*(1), 1-10.

Chiquet, J., Robin, S., & Mariadassou, M. (2019). *Variational inference for sparse network reconstruction from count data.* Paper presented at the International Conference on Machine Learning.

Choi, H., Gim, J., Won, S., Kim, Y. J., Kwon, S., & Park, C. (2017). Network analysis for count data with excess zeros. *BMC genetics, 18*(1), 1-10.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., . . . Bueno, R. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS genetics, 5*(8), e1000602.

Collins, F. S., & Barker, A. D. (2007). Mapping the cancer genome. *Scientific American, 296*(3), 50-57.

Dai, L., Chen, K., Sun, Z., Liu, Z., & Li, G. (2018). Broken adaptive ridge regression and its asymptotic properties. *Journal of multivariate analysis, 168*, 334-351.

de Magalhaes, J. P. (2012). Programmatic features of aging originating in development: aging mechanisms beyond molecular damage? *FASEB J, 26*(12), 4821-4826. doi:10.1096/fj.12-210872

de Magalhaes, J. P., Costa, J., & Church, G. M. (2007). An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J Gerontol A Biol Sci Med Sci, 62*(2), 149-160.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., . . . Estelle, J. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics, 14*(6), 671-683.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis, 90*(1), 196-212.

Dong, X., Milholland, B., & Vijg, J. (2016). Evidence for a limit to human lifespan. *Nature, 538*(7624), 257-259. doi:10.1038/nature19793

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics, 11*(1), 1-9.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics, 9*(3), e1003348.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature methods, 9*(3), 215-216.

Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature protocols, 12*(12), 2478-2492.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*(456), 1348-1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the royal statistical society: series B (statistical methodology), 70*(5), 849-911.

Fei, Z., Zhu, J., Banerjee, M., & Li, Y. (2019). Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. *Biometrics, 75*(2), 551-561.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist, 125*(1), 1-15.

Fernández-Majada, V., Welz, P.-S., Ermolaeva, M. A., Schell, M., Adam, A., Dietlein, F., . . . Schumacher, B. (2016). The tumour suppressor CYLD regulates the p53 DNA damage response. *Nature Communications, 7*(1), 1-14.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science, 303*(5659), 799-805.

Gallopin, M., Rau, A., & Jaffrézic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PloS one, 8*(10), e77503.

Gill, D., Parry, A., Santos, F., Hernando-Herraez, I., Stubbs, T. M., Milagre, I., & Reik, W. (2021). Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *bioRxiv*, 2021.2001.2015.426786. doi:10.1101/2021.01.15.426786

Gorbunova, V., & Seluanov, A. (2009). Coevolution of telomerase activity and body mass in mammals: from mice to beavers. *Mech Ageing Dev, 130*(1-2), 3-9. doi:10.1016/j.mad.2008.02.008

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 326*(1233), 119-157.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., . . . Gao, Y. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell, 49*(2), 359-367.

Harper, J. M., Salmon, A. B., Leiser, S. F., Galecki, A. T., & Miller, R. A. (2007). Skin-derived fibroblasts from long-lived species are resistant to some, but not all, lethal stresses and to the mitochondrial inhibitor rotenone. *Aging Cell, 6*(1), 1-13. doi:https://doi.org/10.1111/j.1474-9726.2006.00255.x

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology, 14*(10), 1-20.

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

Horvath, S., Gurven, M., Levine, M. E., Trumble, B. C., Kaplan, H., Allayee, H., . . . Rickabaugh, T. M. (2016). An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology, 17*(1), 1-23.

Imbert, A., Valsesia, A., Le Gall, C., Armenise, C., Lefebvre, G., Gourraud, P.-A., . . . Villa-Vialaneix, N. (2018). Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics, 34*(10), 1726-1732.

Jusufović, E., Rijavec, M., Keser, D., Korošec, P., Sodja, E., Iljazović, E., . . . Košnik, M. (2012). let-7b and miR-126 are down-regulated in tumor tissue and correlate with microvessel density and survival outcomes in non–small–cell lung cancer.

Kawaguchi, E. S., Shen, J. I., Suchard, M. A., & Li, G. (2021). Scalable algorithms for large competing risks data. *Journal of Computational and Graphical Statistics, 30*(3), 685-693.

Kawaguchi, E. S., Suchard, M. A., Liu, Z., & Li, G. (2020). A surrogate ℓ 0 sparse Cox's regression with applications to sparse high - dimensional massive sample size time - to - event data. *Statistics in medicine, 39*(6), 675-686.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol, 34*(7), 1812-1819. doi:10.1093/molbev/msx116

Lange, K. (2003). *Mathematical and statistical methods for genetic analysis*: Springer Science & Business Media.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics, 9*(1), 1-13.

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., . . . Li, Y. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY), 10*(4), 573.

Li, C., Haghani, A., Robeck, T., Villar, D., Lu, A., Zhang, J., . . . Adams, D. (2021). Epigenetic predictors of maximum lifespan and other life history traits in mammals. *bioRxiv*.

Li, C. Z., Kawaguchi, E. S., & Li, G. (2021). A New ℓ 0-Regularized Log-Linear Poisson Graphical Model with Applications to RNA Sequencing Data. *Journal of Computational Biology, 28*(9), 880-891.

Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics, 13*(3), 523-538.

Li, N., Peng, X., Kawaguchi, E., Suchard, M. A., & Li, G. (2021). A scalable surrogate L0 sparse regression method for generalized linear models with applications to large scale data. *Journal of Statistical Planning and Inference, 213*, 262-281.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems, 23*.

Lowe, R., Barton, C., Jenkins, C. A., Ernst, C., Forman, O., Fernandez-Twinn, D. S., . . . Rakyan, V. K. (2018). Ageing-associated DNA methylation dynamics are a molecular readout of lifespan variation among mammalian species. *Genome Biology, 19*(1), 22. doi:10.1186/s13059-018-1397-1

Lu, A. T., Fei, Z., Haghani, A., Robeck, T. R., Zoller, J. A., Li, C. Z., . . . Horvath, S. (2021). Universal DNA methylation age across mammalian tissues. *bioRxiv*, 2021.2001.2018.426733. doi:10.1101/2021.01.18.426733

Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., . . . Stewart, J. D. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY), 11*(2), 303.

Lu, Y., Brommer, B., Tian, X., Krishnan, A., Meer, M., Wang, C., . . . Sinclair, D. A. (2020). Reprogramming to recover youthful epigenetic information and restore vision. *Nature, 588*(7836), 124-129. doi:10.1038/s41586-020-2975-4

Ma, L., Reinhardt, F., Pan, E., Soutschek, J., Bhat, B., Marcusson, E. G., . . . Weinberg, R. A. (2010). Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nature biotechnology, 28*(4), 341-347.

Mayne, B., Berry, O., Davies, C., Farley, J., & Jarman, S. (2019). A genomic predictor of lifespan in vertebrates. *Scientific Reports, 9*(1), 17866. doi:10.1038/s41598-019-54447-w

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., . . . Bejerano, G. (2010a). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol, 28*(5), 495-501. doi:10.1038/nbt.1630

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., . . . Bejerano, G. (2010b). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol, 28*. doi:10.1038/nbt.1630

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics, 34*(3), 1436-1462.

Mitteldorf, J. (2016). An epigenetic clock controls aging. *Biogerontology, 17*(1), 257-265. doi:10.1007/s10522-015-9617-5

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology, 38*(1), 23-38.

Ocampo, A., Reddy, P., Martinez-Redondo, P., Platero-Luengo, A., Hatanaka, F., Hishida, T., . . . Izpisua Belmonte, J. C. (2016). In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming. *Cell, 167*(7), 1719-1733 e1712. doi:10.1016/j.cell.2016.11.052

Palla, L., & Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics, 97*(2), 250-259.

Peng, J., Mo, R., Ma, J., & Fan, J. (2015). let-7b and let-7c are determinants of intrinsic chemoresistance in renal cell carcinoma. *World journal of surgical oncology, 13*(1), 1-8.

Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H., . . . Valdes, A. M. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research, 20*(4), 434-439.

Rando, Thomas A., & Chang, Howard Y. (2012). Aging, Rejuvenation, and Epigenetic Reprogramming: Resetting the Aging Clock. *Cell, 148*(1–2), 46-57. doi:10.1016/j.cell.2012.01.003

Rodríguez-Matellán, A., Alcazar, N., Hernández, F., Serrano, M., & Ávila, J. (2020). In Vivo Reprogramming Ameliorates Aging Features in Dentate Gyrus Cells and Improves Memory in Mice. *Stem Cell Reports, 15*(5), 1056-1066. doi:10.1016/j.stemcr.2020.09.010

Ronget, V., & Gaillard, J. m. (2020). Assessing ageing patterns for comparative analyses of mortality curves: Going beyond the use of maximum longevity. *Functional Ecology, 34*(1), 65-75.

Saloustros, E., Salpea, P., Qi, C.-F., Gugliotti, L. A., Tsang, K., Liu, S., . . . Stratakis, C. A. (2015). Hematopoietic neoplasms in Prkar2a-deficient mice. *Journal of Experimental & Clinical Cancer Research, 34*(1), 1-8.

Sarkar, T. J., Quarta, M., Mukherjee, S., Colville, A., Paine, P., Doan, L., . . . Sebastiano, V. (2020). Transient non-integrative expression of nuclear reprogramming factors promotes multifaceted amelioration of aging in human cells. *Nature Communications, 11*(1), 1545. doi:10.1038/s41467-020-15174-3

Segrè, A. V., Consortium, D., Investigators, M., Groop, L., Mootha, V. K., Daly, M. J., & Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS genetics, 6*(8), e1001058.

Sen, P., Shah, P. P., Nativio, R., & Berger, S. L. (2016). Epigenetic Mechanisms of Longevity and Aging. *Cell, 166*(4), 822-839. doi:10.1016/j.cell.2016.07.050

Srivastava, S., & Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research, 38*(17), e170-e170.

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell, 126*(4), 663-676. doi:10.1016/j.cell.2006.07.024

Tian, X., Firsanov, D., Zhang, Z., Cheng, Y., Luo, L., Tombline, G., . . . Gorbunova, V. (2019). SIRT6 Is Responsible for More Efficient DNA Double-Strand Break Repair in Long-Lived Species. *Cell, 177*(3), 622-638 e622. doi:10.1016/j.cell.2019.03.043

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal statistical society: series B (Methodological), 58*(1), 267-288.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., . . . Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers, 1*(1), 1-21.

Van de Geer, S., Bühlmann, P., Ritov, Y. a., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of statistics, 42*(3), 1166-1202.

Vaupel, J. W. (2003). Post-darwinian longevity. *Population and Development Review, 29*, 258-269.

Vincze, O., Colchero, F., Lemaître, J.-F., Conde, D. A., Pavard, S., Bieuville, M., . . . Maley, C. C. (2021). Cancer risk across mammals. *Nature*, 1-5.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics, 101*(1), 5-22.

Vu, H., & Ernst, J. (2020). Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *bioRxiv*, 2020.2011.2017.387134. doi:10.1101/2020.11.17.387134

Wan, Y.-W., Allen, G. I., Baker, Y., Yang, E., Ravikumar, P., Anderson, M., & Liu, Z. (2016). XMRF: an R package to fit Markov Networks to high-throughput genetics data. *BMC systems biology, 10*(3), 347-355.

Weber, J., de la Rosa, J., Grove, C. S., Schick, M., Rad, L., Baranov, O., . . . Engleitner, T. (2019). PiggyBac transposon tools for recessive screening identify B-cell lymphoma drivers in mice. *Nature Communications, 10*(1), 1-16.

Wilkinson, G. S., Adams, D. M., Arnold, B. D., Ball, H. C., Breeze, C. E., Carter, G., . . . Horvath, S. (2020). Genome Methylation Predicts Age and Longevity of Bats. *bioRxiv*, 2020.2009.2004.283655. doi:10.1101/2020.09.04.283655

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics, 5*(4), 2493-2518.

Xu, C., & Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association, 109*(507), 1257-1269.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Montgomery, G. W. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics, 42*(7), 565-569.

Yang, J.-H., Griffin, P., Vera, D., Apostolides, J., Hayano, M., Meer, M., . . . Sinclair, D. (2019). Erosion of the Epigenetic Landscape and Loss of Cellular Identity as a Cause of Aging in Mammals. *SSRN Electronic Journal*. doi:10.2139/ssrn.3461780

Yin, J., Bai, Z., Song, J., Yang, Y., Wang, J., Han, W., . . . Yang, Y. (2014). Differential expression of serum miR-126, miR-141 and miR-21 as novel biomarkers for early detection of liver metastasis in colorectal cancer. *Chinese Journal of Cancer Research, 26*(1), 95.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics, 38*(2), 894-942.

Zhao, H., Sun, D., Li, G., & Sun, J. (2018). Variable selection for recurrent event data with broken adaptive ridge regression. *Canadian Journal of Statistics, 46*(3), 416-428.

Zhao, H., Wu, Q., Li, G., & Sun, J. (2019). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*.

Zhao, S. D., & Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis, 105*(1), 397-411.

Zhou, W., Triche, T. J., Jr, Laird, P. W., & Shen, H. (2018). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research, 46*(20), e123-e123. doi:10.1093/nar/gky691

Zhu, Y., Zhang, C., Gu, C., Li, Q., & Wu, N. (2016). Function of deubiquitinating enzyme USP14 as oncogene in different types of cancer. *Cellular Physiology and Biochemistry*, *38*(3), 993-1002.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical Association*, *101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.