Restriction of low variance features for the construction of more accurate epigenetic clocks

By MBI candidate: Vinson Zeng

Final report for BMIF 898 (Master's Project)

Supervisor: Dr. Robert Gooding

Queen's University

Kingston, Ontario, Canada

July 8, 2022

Abstract

Epigenetic clocks are predictors of age based on the DNA methylation levels of specific CpG sites in various tissues and cells. These clocks are typically used to estimate the biological or DNA methylation age of organisms, which may differ from their chronological age, but have been used to estimate mortality as well. There has been a growing interest in these clocks over the past decade and, with it, a need for more accurate and efficient clock construction methods. A penalized regression model, specifically elastic net regularization, is the most commonly used method to create clocks. The models regress chronological age on measured CpG probes. However, few studies employ a discrete feature selection method and rely solely on the elastic net for both feature selection and prediction. Utilizing a feature selection approach potentially increases the efficiency of clock development by reducing the number of CpG sites input into elastic net regression. Clocks developed using a feature selection method may also prove to be more accurate as some features may be irrelevant data, or noise, which can decrease accuracy and prediction results in statistical learning models. In our study, we apply the classic feature selection method of excluding low variance features to the development of epigenetic clocks using elastic net regression. We also investigate the effects of excluding high variance features and using subsets of features ranging from low to high variances for clock development. The feature selection methods are applied to a dataset consisting of 256801 CpG features from a merged TCGA collection of BRCA, KIRC, KIRP, LUAD, LUSC, and THCA normal samples (*n*=431). We construct clocks to predict age and find that clocks utilizing higher variance CpG probes have better predictive accuracy. The best performing clock uses 213 CpG sites with an external test RMSE of 6.165854 (*n*=121), compared to a RMSE of 6.457379 from a clock constructed without feature selection (All-features clock) which uses 252 CpG sites. Furthermore, we evaluate and compare both these clocks on cancer datasets for TCGA KIRC, KIRP, and BRCA. Our findings indicate that the All-features clock predicts more epigenetic age acceleration, the difference between DNA methylation age and chronological age, for cancer tissue. Excluding low variance features may result in a simpler and more accurate epigenetic clock for predicting biological age but, depending on the variance threshold, the prediction of age acceleration may be impacted.

Table of Contents

1 Introduction

Epigenetics is the study of heritable changes in gene function. These changes are attributed to modifications in gene expression and activity. The modifications are not encoded in DNA sequences but instead they involve alterations in how DNA is read and translated into proteins. Epigenetic changes occur naturally with age[1], as a part of normal development[2], in response to lifestyle factors such as diet[3,4] and exercise[5–8], and in response to environmental exposures such as social experiences and pollutants[4,8,9]. Epigenetics is responsible for phenotypic differences between organisms with identical DNA. Most cells in an organism contain the same DNA, but they vary in types and functions due to qualitative and quantitative differences in their gene expression[10].

The major mechanisms responsible for epigenetic changes include DNA methylation (DNAm), histone modification, and non-coding RNA. DNAm is the covalent addition of a methyl group to DNA at a cytosine nucleotide. DNAm in humans generally occurs at sites where cytosine precedes a guanine nucleotide, known as CpG sites (CpGs)[10]. High density CpG areas are known as CpG islands. These islands are typically associated with gene promoters, which are sequences of DNA involved in initiating transcription. There are also 2 kilobase (kb) flanking regions known as CpG island shores, CpG island shelves which are 2-4 kb from CpG islands, and differentially methylated regions amongst other methylation sites of interest[11]. DNAm regulates gene expression by inhibiting the binding of transcription factors or by recruiting proteins involved in gene repression. A family of DNA methyltransferases (Dnmts) catalyze DNAm by transferring a methyl group from *S*-adenyl methionine to the fifth carbon of a cytosine residue. DNAm involves the enzymes Dnmt1, Dnmt3a, and Dnmt3b, which directly catalyze the addition of methyl groups onto DNA. DNA demethylation occurs passively in dividing cells through inhibition or dysfunction of Dnmt1, or actively through enzymatic reactions to revert the 5-methylcytosine (5mC) back to a naked cytosine[12]. Histones are spool-like proteins that DNA coils around to form condensed units. Histone modification involves the addition or subtraction or specific chemical groups, which can modify the conformation of the DNA packaging structure. This plays a role in gene regulation by enabling or disabling access for specific proteins to read the genes[13]. Non-coding RNA can regulate gene expression by interacting with DNA, RNA, and proteins. They can contribute to the breakdown of coding RNA, alter chromatin function and structure, and affect the transcription of genes, RNA splicing, and translation[14–16]. In a general sense, these mechanisms regulate gene expression by turning genes on and off.

From the epigenetic mechanisms described, DNAm has emerged as an attractive biomarker due to its accessibility for quantitative measurements. Advancements in methylation array technology has enabled high-throughput, genome-wide methylation profiling which provide high-resolution data for research[11,17]. Patterns of DNAm changes across the genome have been previously associated with aging[18–22] and advancements in DNAm array technologies have facilitated the development of epigenetic clocks built using CpGs found to change with age[19,23–26]. Specifically, more advanced array technology introduces more target CpG sites. For example, the Illumina DNAm arrays have improved from being able to analyze over 27000 CpG sites with its Infinium HumanMethylation27 Beadchip (27K array)[17], followed by the 450K array with over 480000 CpG sites[11], then over 850000 CpG sites with the EPIC array[27]. Epigenetic clocks can measure biological age, the age an individual appears to be based on genetics and environmental factors, as opposed to chronological age which is measured in actual time.

Biological age Is generally measured by using statistical learning methods which regress chronological age on selected CpGs. In addition to measuring age, there are epigenetic clocks which can be used to predict disease and mortality risk[25,26,28]. The CpGs with the most significant hyper- or hypomethylation correlation with age are selected and usually weighted in a linear model. To measure the methylation status of CpG sites, the Illumina arrays may utilize one or both of two types of DNA analysis chemistries the Infinium I and Infinium II (Figure 1). The Infinium I assay employs two beads which correspond to methylated (C base) and unmethylated (T base) states for each CpG locus[29]. Incorporation of fluorescently labeled nucleotides during single-base extension at each probe type indicates the methylation or percentage of a particular base level within a sample[30]. The Infinium II assay employs a single bead at for each CpG locus in which a labeled G or A base is added during single base extension to the complimentary methylated C or unmethylated T. Detection of the intensity of fluorescence is calculated as a beta value, a continuous variable between 0 and 1, which is a ratio of methylated to intensity values from the combined assays. This ratio is used to determine the percentage of methylation[29].

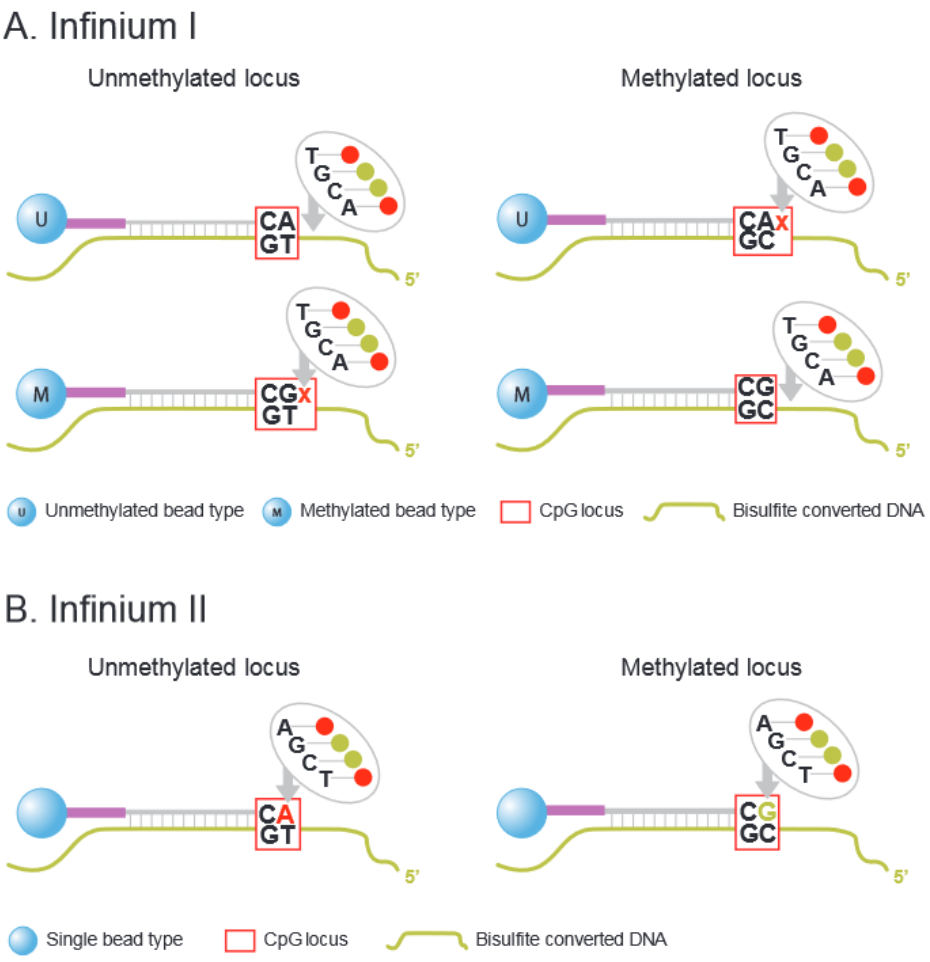

Figure 1. Infinium I uses 2 bead types per CpG locus. Infinium II uses 1 type per CpG locus with the methylation state determined post-hybridization[30].

Horvath's clock

The Horvath clock, published in 2013, was the first multi-tissue predictor of age created. It used target-specific probes common to two DNAm analysis arrays, the Illumina 27K and 450K platforms[23]. The platforms can provide quantitative measurements for over 27000 and 480000 CpG sites, respectively, within a given DNA sample[11,17]. 7844 healthy, non-cancerous, samples were collected from 82 datasets, across 51 different tissue and cell types. A training dataset, derived from the collective data, was used to define the age predictor. A modified version of chronological age was regressed onto 21369 CpG probes in a penalized regression model (elastic net)[31]. This statistical model automatically selected 353 CpGs. The selected probes were found to correlate with age and the weighted averages of their statistical measures formed the epigenetic clock, which has a mean error of 3.6 years. The predicted age using Horvath's clock is defined as DNAm age. The clock allows for the comparison of ages across different areas of the body within the same individual, without the need for any adjustments to the model. This may help to identify evidence of disease related epigenetic age acceleration (EAA), which is the difference between DNAm age and chronological age. Horvath hypothesized that the DNAm age measures the cumulative work done by an epigenetic maintenance system (EMS), which theoretically helps to maintain epigenetic stability[23]. Although observations consistent with this hypothesis are observed in some cancers, the mechanisms of the EMS require more conclusive research[32].

Epigenetic clocks released before Horvath's clock publication

The first recognized epigenetic clock, the Bocklandt Clock, was created in 2011 based on DNA extracted from saliva samples from 34 male identical twin pairs, for a total of 68 samples[19]. This model was used to predict the age of a subject through saliva with a mean absolute difference between the observed and predicted age, or mean error, of 5.2 years. There were 88 probes originally identified which correlated with age but, after validation using a sample of unrelated males and females, a model explaining 73% of the variance in age was built using just two CpGs. Later in 2011 Koch and Wagner developed a clock, the Epigenetic-Aging-Signature, using 5 CpGs[33]. This clock predicted the age in the multiple cell types [list cell types?] used to train and validate the model with a mean error of 11 years. Koch, Wagner, and their team developed a clock using 6 CpGs in the following year of 2012 which could predict the state of human cellular senescence, or cell passage number, in fibroblasts and mesenchymal stem cells[34]. The average difference between actual and predicted passage number was three, with a standard deviation of four. In 2013 the Hannum clock made its debut at the beginning of the year before Horvath's clock and was the first clock developed which used blood[24]. Hannum's clock is composed of 71 CpGs and has a mean error of 3.9 years. Hannum's clock was also representative of other tissues, but this required an adjustment of the model for each specific tissue to achieve a comparable error to that found in blood. These four clocks gave an early glimpse into the potential utility of DNA-m based models, such as use in forensics, routine medical screening, and precision medicine[19,24,33,34]. In addition, more accurate age estimators appear to be associated with a greater number of CpGs as observed from the few early clocks described.

Epigenetic clocks released after Horvath's clock publication

Since the release of Horvath's clock in 2013, there has been at least one epigenetic clock released each year[35–37]. In 2014, there were at least three different epigenetic clocks released. Weidner et al[38] released a clock for blood which consisted of only 3 CpGs and had an average error of 4.5 years. The model started with 102 CpGs with an average error of 3.34 years and was then reduced to 99 CpGs with an average error of 4.12 years after validating the model on Hannum's dataset, which was based on the

Illumina 450K array. The model was then restricted to the most relevant CpGs by recursive feature elimination to facilitate site-specific analysis of DNAm, as opposed to profiling approaches. This resulted in the three CpGs for a minimized clock, which utilizes fewer CpGs. Florath et al[39] built a clock for age prediction using blood based on 17 CpGs, with an average error of 2.6 years. However, one drawback of the clock was that the cohort study used was narrow as the ages ranged between 50 and 75 years. Polanowski et al[40] developed a clock using three CpG sites for humpback whales based on skin samples, with an average error of 2.991 years. This clock was the first epigenetic age estimator for animals.

In 2015, Huang et al[41] published a clock with an average error of 7.87 years for age prediction in blood and bloodstain samples. The study was conducted in a Chinese Han population. Five significantly age-correlated CpG sites were identified using pyrosequencing and four were used in the final model after a stepwise regression analysis. Zbiec-Piekarska et al[42] published a clock in the same year for age prediction in blood that was conducted in a Polish population. This age prediction model was based on DNAm in 5 CpG sites and had an average error of 3.9 years. Both Huang and Zbiec-Piekarska clocks were developed with a focus on forensic application, hence the models aimed to use as few CpGs as possible for a more cost-effective DNAm profiling approach and for the potential to be upscaled.

In 2016, Yang et al.[43] released a mitotic clock, which measures the total number of lifetime cell divisions, called epiTOC (Epigenetic Timer of Cancer) and Knight et al.[44] released an epigenetic clock for gestational age at birth. The epiTOC study focused on promoter CpG sites based on blood samples that were localized to Polycomb group target genes. Polycomb proteins are involved in gene silencing and regulate histone methylation profiles of various genes responsible for controlling cellular pathways[45]. These sites are umethylated in fetal tissues and their degree of methylation correlates with the increased rate of cell division during aging. The model was built using 385 CpGs and tracks cell divisions[43]. In addition, its tick rate is accelerated in cancer, premalignant epithelial lesions, and in normal epithelial cells which have been exposed to major carcinogens. The gestational age clock was built using 148 CpG sites based on blood methylation data[44]. It was determined that an accurate prediction could be made for gestational age between 24 and 44 weeks with an average error of 1.49 weeks. Accelerated gestational age may reflect developmental maturity of a neonate, epigenetic programming resulting from early life environmental exposures such as pregnancy disorder or prenatal stress, or the difference between DNAm gestational age and chronological age may reflect the variable nature of gestational age estimations[44].

In 2017, we begin to see an uptick in epigenetic clocks relative to previous years. Zhang et al[46] released a model which uses DNAm signatures in blood. It yields a mortality risk score based on 10 CpGs which demonstrates a strong association with all-cause mortality, cardiovascular disease, and cancer mortality. It was the first study to link DNAm across various genes to mortality in the general population[46]. Cho et al[47] released a minimized clock for age prediction in blood with an average error of 4.2 years. It is based on 5 CpGs which are located near the same genes examined in the Zbiec-Piekarska et al[42] study from 2015. Several non-human epigenetic clocks were released during this year as well. Wang et al[48], Petkovich et al[49], and Stubbs et al[50] released age predictors for mice and are based on 107, 90, and 329 CpGs, respectively. Thompson et al[51] released an epigenetic clock for dogs and wolves based on 115 CpGs.

In 2018 several animal clocks, another clock involving Horvath, a mitotic clock, and a composite clock were released. The non-human clocks include a model for chimpanzees[52] and 5 models for mice[53,54].

Horvath et al[55] released an epigenetic clock for skin and blood based on 391 CpGs, referred to as the Skin and Blood clock. It predicts the age of fibroblasts, keratinocytes, endothelial cells, buccal cells, lymphoblastoid cells, skin, blood, and saliva better than Horvath's original clock. The Skin and Blood Clock has a root mean square error (RMSE) of 7.64 and Horvath Clock has a RMSE of 15.74 for the mentioned sample types[55]. MiAge, is a DNAm-based mitotic age calculator developed by Youn et al[56] which is based on the MiAge model, a novel statistical framework. It is based on 268 CpGs and the clock shows accelerated mitotic age across 13 different cancer types, with accelerated mitotic age associated with worse cancer survival. Levine et al[25] developed a composite age estimator, which uses CpGs correlated with physiological aging, by averaging the weight of 10 clinical characteristics. Chronological age, creatinine, glucose levels, C-reactive protein levels, albumin, lymphocyte percentage, mean cell volume, red blood cell distribution width, white blood cell count, and alkaline phosphatase were regressed on DNAm levels in blood using a penalized regression model. The resulting 513 CpGs formed the DNAm PhenoAge clock, which can predict all-cause mortality, health span, physical functioning, and cancer better than the first generation of DNAm age estimators[25,32].

In 2019, clocks released include a composite clock called GrimAge by Lu et al[26], the Zhang clock by Zhang et al[28], a DNAm telomere length (DNAmTL) estimator by Lu, Seeboth, et al[57], the Pediatric-Buccal-Epigenetic (PedBE) clock by McEwen et al[58], and a minimized clock by Jung et al[59]. The GrimAge clock uses a composite measure based on the DNAm levels of CpGs associated with seven plasma proteins, chronological age, sex, and smoking pack years to predict lifespan and healthspan[26]. The plasma proteins and smoking pack years act as surrogate DNAm biomarkers of stress factors and physiological risk factors. A large-scale meta-analysis revealed that DNAm GrimAge outperforms all prior DNAm-based predictors for lifespan prediction[26]. The Zhang clock, trained on 13661 samples (259 from saliva and 13,402 from blood), is based on 514 CpGs. It outperforms both Hannum and Horvath clocks in estimating blood age with a RMSE of 2.04 years[28]. Furthermore, the Zhang clock study suggests that a sufficiently large training set could result in a near-perfect chronological age predictor. The DNAmTL estimator is based on 140 CpGs which resulted from the regression of leukocyte telomere length against blood methylation data[57]. DNAmTL rivals Hannum's and Horvath's clock but is outperformed by PhenoAge and GrimAge. However, DNAmTL outperforms typical measures of leukocyte telomere length (LTL), such as quantitative polymerase chain reaction[60], in predicting time-to-death, time-to-coronary heart disease, time-to-congestive heart failure, and association with smoking history[57]. In addition, DNAmTL is more strongly associated with age than measured LTL. The PedBE clock was developed to address the lack of accurate predictions from epigenetic clocks in the pediatric age range[58]. This is due to the higher rate of DNAm changes in that age range, described as a high tick rate during organismal development by Horvath[23]. The PedBE clock uses 1032 blood, buccal cells, and saliva training samples from individuals aged 0 to 20 years old and it is based on 94 CpGs[58]. The minimized clock by Jung et al[59] is based on 5 CpGs and has an RMSE of 5. The model is trained on 300 samples, with 100 from blood, 100 from saliva, and 100 from buccal swabs. Like previous minimized clocks, it would be useful in forensic analysis due to its ability to be applied to a broad spectrum of tissues and body fluids with an adequate predictive accuracy.

In 2020, published clocks include epigenetic clocks for various animals[61–70], a minimized clock by Dias et al[71], a skin clock by Boroni et al[72], a skeletal muscle clock by Voisin et al[73], and the DunedinPoAm Clock by Belsky et al[74]. The animals for which clocks were created for in this year include mice[61], dogs[62], chimpanzees[63], naked mole rats[64], seabass[65], zebrafish[66], rats[67,68], prairie voles[69], and deer[70]. The

minimized clock by Dias et al[71] uses blood samples from deceased individuals to estimate DNAm age. It is trained on 51 samples and uses 5 CpGs. The deceased-blood DNAm age predictor model has a 4-fold cross-validation RMSE of 7.43 and an independent test (*n*=19) RMSE of 10.98. The skin clock by Boroni et al[72] uses 2266 CpGs, one of the largest number of clock CpGs observed, and is trained on 249 skin samples[72]. It has a test RMSE of 4.98, which is better than Horvath's Clock and the Skin and Blood Clock for DNAm age prediction for skin (RMSEs of 15.74 and 7.64, respectively)[23,55]. The skeletal muscle clock is trained on 682 skeletal muscle samples and is based on 200 CpGs[73]. The clock has a median error of 4.6 years, which outperforms Horvath's pan-tissue clock (median error of 13.31 years) for DNAm age estimation in skeletal muscle. The skeletal muscle clock may be helpful in investigating the impact of exercise and diet on muscle-specific biological aging processes. The DunedinPoAm Clock is a blood-based DNAm predictor of the pace of biological aging for an individual at a single point in time[74]. It uses a panel of 18 blood-chemistry and organ-system-function biomarkers from a cohort with the same birth year and place. The clock is trained on 810 blood samples and is based on 46 CpGs.

In 2021, animal clocks released included those for velvet monkeys[75,76], pigs[77], elephants[78], cats[79], macaques[75,80], marmosets[75,81], baboons[75], bats[82], beluga whales[83], sheep[84], deer mice[85], and cattle[86]. The fetal brain clock (FBC)[87], the NEOage clocks[88], and DeepMAge[89], the first deep learning DNAm aging clock, were also released in the same year. The FBC is trained on human prenatal brain samples and is used to investigate the epigenetic age of induced pluripotent stem cells (iPSCs) and differentiated neurons (iPSC-neurons) during the earliest stages of human neurodevelopment[87]. The FBC was found to perform better than both Horvath's clock[23] and the gestational age clock from 2016[44] at predicting age in fetal brain samples. Based on the results, the FBC is suggested to be the best available tool to profile neuronal development models. The NEOage clocks consist of four separate models, each used to predict post-menstrual and postnatal age[88]. They are based on buccal cells from 542 preterm infants and utilize 303 to 522 CpGs with varying overlap between the clocks. DeepMAge is a novel neural network regressor which is based on 1000 blood samples from 17 studies[89]. It outperforms Horvath's clock in seven out of 15 datasets using both median absolute error and Pearson's r, and in 13 out of 15 studies according to at least one metric. Furthermore, DeepMAge is slightly more accurate than both Horvath and Hannum clocks at predicting the age of healthy individuals[89]. However, Horvath's model is a pan-tissue predictor while DeepMAge is only trained in blood DNAm data.

In the current year of 2022, animal epigenetic clocks published include ones for yellow-bellied marmots[90], naked mole-rats[91], horses[92], roe deer[93], mouse cells in culture[94], and the first epigenetic clock for a reptile (green turtles)[95]. Human epigenetic clocks which have been published to this point include Dunedin pace of aging (DunedinPACE)[96], a monocyte-based DNAm clock (monoDNAmAGE)[97], and a pan-tissue DNAm clock based on deep learning (AltumAge)[98]. DunedinPACE is a blood biomarker of the pace of aging which addresses the limited biological scope of the previously described DunedinPoAm clock from 2020[74,96]. DunedinPoAm was based on a twelve-year cohort study and had limited precision. DunedinPACE incorporates new data from the Dunedin cohort to extend the observable scope of biological change over time and restricts DNAm probes with inconsistent measurability[99] to improve the power and precision of the model overall[96]. MonoDNAmAGE is based on 186 CpG sites from monocyte methylomes and was developed to assess the impact of alcohol consumption on monocyte age[97]. The results suggest a nonlinear effect of alcohol consumption on biological aging, with nonheavy use decreasing and heavy consumption increasing epigenetic age. AltumAge is a deep neural network which uses 20318 CpG sites for pan-tissue age prediction, with the

samples gathered from 142 datasets[98]. The results demonstrate that AltumAge predicts higher age acceleration than Horvath's clock for tumors and for cells exhibiting age-related changes in vitro, such as samples from patients with type 2 diabetes. Deep learning epigenetic clocks[89,98] have had promising results so far which appear to outperform the common approach of regularized linear regression.

Applications of epigenetic clocks

A positive or negative EAA value suggests that the tissue or cells being investigated are aging faster or slower, respectively, than expected. Epigenetic clocks have been used in age acceleration studies to investigate the impact of various factors such as genetics, diet, lifestyle, environment, illness, medication, and disease therapy on EAA in various cells, tissue, and organs. Epigenetic aging studies include, but are not limited to: COVID-19[100–102], postpartum sleep loss[103], early life trauma[104], maltreatment in children[105,106], risk of atrial fibrillation[107], diet and lifestyle interventions[108–110], spaceflight effects[111], schizophrenia[112], autism spectrum disorders[113], exposure to ionizing radiation[114], radiation therapy in cancer[115], dementia risk[116], estrogen exposure on breast epigenetic age[117], work-related trauma[118], HIV in adolescents[119], antiretroviral therapy in adults with HIV[120], age-related macular degeneration[121], aplastic anemia[122], age-related hearing loss[123], cardiovascular health in school-aged children[124] and adults[125], impact of neighbourhood social environments[126], adult survivors of childhood cancer[127], effects of socioeconomic status[128], effects of age, sex, education, and ethnicity[129], exercise effects on skeletal muscle epigenetic aging[130,131], depression[132], exposure to carcinogens[133], common medications[134], hepatitis[135], renal function[136], Williams syndrome[137], cigarette smoking[138], vitamin D levels[139], and allografts[140,141].

Despite the success of recent epigenetic clocks utilizing deep learning methods[89,98], approaches such as regularized regression models are preferrable for applications in computational biology for model interpretability and since data are less abundant[142]. From the *glmnet* package[143], elastic net uses the least angle regression algorithm to estimate parameters[31,144,145] and it combines the shrinkage-type regression methods of least absolute shrinkage and selection operator (also known as LASSO; L1 regularization)[146] and Ridge regression (L2 regularization)[147]. Regularized regression models penalize large coefficients, which result in models which are less complex and more interpretable. Using a convex combination of L1 and L2 penalties to handle sparsity and correlated features, the elastic net is well suited for prediction tasks with a greater number of potentially correlated features (p) than samples (n). Datasets which contain many more features than samples ($n \ll p$) are also referred to as high-dimensional data[148]. The elastic net outperforms methods such as LASSO and supervised principal components for large-scale DNAm profiling[149] and it has emerged as one of the most popular methods used to find age associated CpGs (Table 1). However, few epigenetic clock studies utilize a discrete feature selection method and rely exclusively on the elastic net to select CpGs. The addition of a separate feature selection method decreases the number of CpG sites used in the elastic net, which may increase the efficiency of clock development and help to generate a simpler model with more interpretability. One strategy for feature selection is to exclude features with low or near zero variance. Features which are constant or approximately constant across all samples have zero or low variance, respectively, and will not improve the performance of a model. Such features may be irrelevant data termed as attribute noise which can decrease accuracy and prediction results in statistical learning models[150].

In this study, we aimed to characterize the effect of restricting attention to both low and high variance probes on an epigenetic clock. This clock was developed using regularized linear regression (elastic net) for feature selection and prediction. We based our clock on normal tissue samples, which are samples taken adjacent to tumours, from The Cancer Genome Atlas (TCGA) program[151] based on the Illumina 450K platform. The datasets we use in our study are publicly accessible and consist of the following TCGA collections: kidney renal clear cell carcinoma (KIRK), cervical kidney renal papillary cell carcinoma (KIRP), breast invasive carcinoma (BRCA), lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), thyroid cancer (THCA), head-neck squamous cell carcinoma (HNSC), prostate adenocarcinoma (PRAD), liver hepatocellular carcinoma (LIHC), uterine corpus endometrial carcinoma (UCEC), and colon adenocarcinoma (COAD). We evaluate the accuracy and generalizability on the following datasets: GSE101961[152], GSE59157[153], TCGA prostate adenocarcinoma (PRAD), and TCGA liver hepatocellular carcinoma (LIHC). In addition, we evaluate our clock on TCGA KIRP, KIRC, and BRCA tumor samples.

Clock/Study	Cell/tissue types	No. of training samples	Method used to select CpGs	No. CpGs used in model	Application of clock
Bocklandt et al., 2011	Saliva	34 pairs of identical twins (*n* = 68)	Probes with both q-values <0.05 and correlation values >0.57	2	DNAm age
Koch and Wagner, 2011	Fibroblasts/dermis, keratinocytes/epidermis, epithelial cells/cervical smear, $CD4^+$ T-cells and $CD14^+$ monocytes/blood	130	Pavlidis Template Matching and correlation values >0.6	5	DNAm age
Koch et al., 2012	Fibroblasts, mesenchymal stem cells	51	Pavlidis Template Matching	6	Predict state of human cellular senescence (passage number)
Hannum et al., 2013	Blood	482	Elastic net	71	DNAm age
Horvath's Clock/Horvath et al., 2013	51 different cell/tissue types	3931	Elastic net	353	DNAm age
Weidner et al., 2014	Blood	575	Pearson correlation (either r > 0.85 or r < -0.85) then recursive feature elimination	3	DNAm age
Florath et al., 2014	Blood	249	Forward selection by minimization of Akaike's Information Criterion (AIC)	17	DNAm age
Huang et al., 2015	Blood	89	CpGs from previous studies[ref], pyrosequencing then stepwise regression	4	DNAm age for forensics

Table 1. Summary of various epigenetic clocks based on human DNAm arrays, organized by ascending publication year.

Clock/Study	Cell/tissue types	No. of training samples	Method used to select CpGs	No. CpGs used in model	Application of clock
Zbiec-Piekarska et al., 2015	Blood	420	41 CpGs selected from Hannum et al. (2013) followed by multivariate linear regression	5	DNAm age for forensics
epiTOC/ Yang et al., 2016	Blood	656	CpGs with false discovery rate threshold (q) < 0.05, then filtered for those mapping within 200 base pairs of a transcription start site	385	Mitotic age, cancer risk
Gestational age clock / Knight et al., 2016	Blood	171	Elastic net	148	Gestational age at birth
Zhang Mortaility Clock; Zhang et al., 2017	Blood	548	LASSO Cox regression	10	DNAm-based mortality risk score
Cho et al., 2017	Blood	100	32 CpGs from Zbiec-Piekarska (ref), multivariate regression	5	DNAm age
MiAge/Youn and Wang, 2018	TCGA methylation data (BRCA, COAD, HNSC, KIRP, LIHC, PRAD, THCA, UCEC)	4020	Novel statistical framework – MiAge model	268	Mitotic age, cancer risk, survival
PhenoAge/Levine et al., 2018	Blood	9926	Elastic net	513	All-cause mortality
Skin and Blood/Horvath et al., 2018	Fibroblasts, keratinocytes, buccal cells, endothelial cells, lymphoblastoid cells, skin, blood, saliva	896	Elastic net	391	DNAm age, improvement over original Horvath clock for various cells
GrimAge/Lu et al., 2019	Blood	1731	Elastic net	1030	All-cause mortality
Zhang clock/Zhang et al., 2019	Blood, saliva	13661	Elastic net	514	DNAm age
Lu et al., 2019	Blood	2256	Elastic net	140	DNAm-based telomere length estimator
DunedinPoAm algorithm/Belsky et al., 2020	Blood	810	Elastic net	46	DNAm-based measure of pace of biological aging
Boroni et al., 2020	Skin	249	Elastic net	2266	Skin DNAm age
Voisin et al., 2020	Skeletal muscle	682	Elastic net	200	Skeletal muscle DNAm age
Fetal brain clock (FBC); Steg et al., 2021	Fetal brain samples	193	Elastic net	107	Prenatal age
NEOage clocks (four epigenetic clocks)/Graw et al., 2021	Buccal cells	542	Elastic net	303-522	Post-menstrual age and post-natal age in preterm infants

Table 1 - Continued (1).

Clock/Study	Cell/tissue types	No. of training samples	Method used to select CpGs	No. CpGs used in model	Application of clock
DeepMAge/Galkin et al., 2021	Blood	4930	Regression via feed-forward neural networks with >3 hidden layers	1000	DNAm age
MonoDNAmAge/Liang et al., 2022	CD14+ monocytes	721	Elastic net	186	Assess impact of alcohol consumption on monocyte DNAm age
DunedinPACE/Belsky et al., 2022	Blood	1037	Elastic net	173	Blood biomarker for pace of aging
AltumAge/Camillo et al., 2022	Multiple tissues from 142 datasets	20318	Neural network with 5 hidden layers	20318	DNAm age

Table 1 - Continued (2).

2 Methods

Data

The datasets utilized in this study are from the TCGA projects on the Genomic Data Commons (GDC) data portal for DNAm data, the Gene Expression Omnibus (GEO) database, and the cBio Cancer Genomics Portal (cBioPortal). TCGA is a collaborative project containing over 2.5 petabytes of genomic data. The GDC Data Portal is a platform that allows users to examine and retrieve cancer data from TCGA. Since 2006, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) have facilitated cancer research by characterizing 33 cancer types. The data is publicly available and includes over 20000 biospecimens[154]. TCGA data we retrieved is from the harmonized database, which uses the latest human reference genome GRCh38 (hg38) and is produced by using a standardized analysis pipeline to process the data[155]. Data harmonization brings together the raw data from various cancer projects and provides a comparable view of information across the different sources. This allows for a highly reliable and standardized workflow for bioinformatics and cancer research. Without harmonization, minor differences in data processing and analysis pipelines across projects may hinder research progress. The GDC generates derived analysis data from submitted FASTQ and BAM files[156]. Sequence data is aligned or realigned to the most recent human genome reference before it is processed into derived data. Array data is processed with suitable methods depending on the data type. Experts in the field of cancer genomics are consulted in the development and maintenance of data processing pipelines. The GEO database is an international, public functional genomics repository which accepts raw and processed data for high-throughput studies of gene expression and genomics[157]. It was created in 2000 and provides access to tens of thousands of studies. The cBioPortal is an open platform resource for the exploration of multidimensional cancer genomics data[158]. It provides integrated access to more than 5000 tumour samples from 20 cancer studies[159]. Datasets used in this study are from TCGA projects KIRC, KIRP, BRCA, LUAD, LUSC, THCA, HNSC, PRAD, LIHC, UCEC, and COAD. The main testing datasets are from GEO under the accession codes GSE101961 and GSE59157. Cancer testing datasets are from TCGA collections KIRC, KIRP, and BRCA. All analysis was done in RStudio[160] .All R code outlining our methods can be found on github (https://github.com/v-zeng/BMIF898_project).

MacIntyre filtered probes

The *keep probes* or *good probes* are the result of probe filtering for various genomic factors. These factors may compromise the ability of the 450K array to accurately measure methylation. MacIntyre et al.[161] compared beta-values from the 450K bead array with whole genome bisulfite sequencing (WGBS) to determine which CpG probes provide accurate or noisy signals. They derived a set of high-quality probes that provide pure measurements of DNA methylation. The correlation of beta-values increased when comparing the high-quality probes only between the 450K and WGBS. This suggests that the remaining probes are affected by some type of genomic factors other than methylation. Probes which hybridize multiple genomic locations make it difficult to determine which genomic regions are responsible for the observed methylation state. Probes which hybridize repetitive regions may introduce confounding factors. Probes which hybridize regions containing small insertions and deletions (INDELs) may affect probe hybridization. Probes which hybridize regions containing single nucleotide polymorphisms (SNPs) may or may not affect probe hybridization. Probes which hybridize regions affected by unknown factors result in large discrepancies between 450K and WGBS beta-values and are not kept. By selecting high-quality probes, the risk of false discovery is reduced.

Data preprocessing

Using TCGAbiolinks[162] in R[163], a package designed for integrative analysis with GDC data, we downloaded 653 normal samples in total from the Illumina 450K platform for the KIRC, KIRP, BRCA, LUAD, LUSC, THCA, HNSC, PRAD, LIHC, UCEC, and COAD datasets. The beta values of 485577 probes of all 653 normal samples were taken. The Macintyre study[161] identified 284482 high-quality probes unaffected by any genomic factors, which were selected for in each of our normal sample datasets. From the Macintyre selected probes, we then removed those with any missing (NA) values. Table 2 provides a summary of the sample count and probe count after omitting NA values for each of the 11 TCGA projects.

Project	Tissue site	Probe count	Sample count
KIRC	Kidney	267461	160
BRCA	Breast	266142	96
THCA	Thyroid	268810	56
HNSC	Head and Neck	267059	50
PRAD	Prostate	267595	50
LIHC	Liver	265455	50
KIRP	Kidney	267388	45
LUSC	Lung	268858	42
COAD	Colon	266186	38
UCEC	Uterus	268242	34
LUAD	Lung	268442	32

Table 2. DNAm probe count from each normal tissue sample dataset. Only probes matching the MacIntyre keep probes and with zero NA values are selected. Tissues are sampled from a total of nine different sites.

Next, we kept only the common probes between all 11 projects, which was determined to be 256801. To measure the spread of the beta value distributions, we used the standard deviation (SD) function in R to compute the SD for each probe. The SD is a measure of the distribution of the values, in which a higher measure indicates a wider spread of values, and a lower measure indicates a closer spread of values. A column was added for the SD and the probes were then sorted in order of ascending SD. Clinical data was retrieved from cBioPortal using the cBioPortalData R package[164]. The patient IDs were used to match the clinical data to the samples, which were then merged with the beta values and

written to a comma-separated values (CSV) file for each dataset. TCGA data preprocessing is outlined in Figure 2. After preprocessing each TCGA dataset individually we arbitrarily selected four tissue types to use in our analysis. All breast, kidney, lung, and thyroid samples, which consists of the BRCA, KIRC, KIRP, LUAD, LUSC, and THCA collections, were merged into a single dataset (*n*=431).

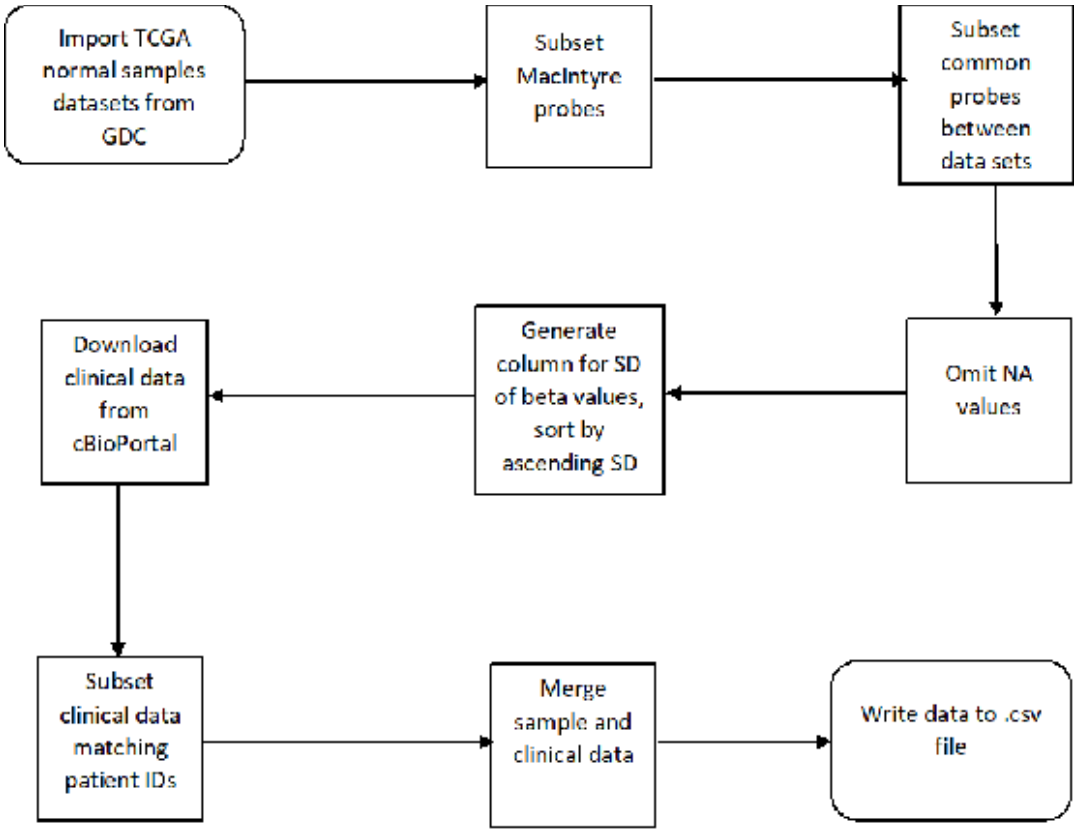

Figure 2. General workflow for TCGA data preprocessing. Normal samples were downloaded from GDC using TCGAbiolinks in R. The 485577 beta values were taken and subset by matching the 284482 MacIntyre good probes and then subset again by the 256801 common probes between the datasets. NA values were then omitted, and the SD was calculated for each feature (probe beta values). Features were sorted in ascending SD order. Clinical data was downloaded using the cBioPortalData R package and then matched to sample patient IDs. Beta values and clinical data were then merged and saved to a DataFrame.

Cross-validation and overall approach

The merged dataset was split into a training and testing set of 75% and 25% of the samples, respectively. Training data was then split into *k* groups or *folds* of approximately equal size for *k-fold* cross-validation (CV). This CV approach utilizes the first fold as a validation set, with the statistical learning method fit on the remaining *k*-1 folds. The process is repeated *k* times, and the test error is computed for a different fold used as the validation set each time. The procedure results in *k* test error estimates which are then averaged for the *k*-fold CV estimate[148]. For our study we used the *glmnet* package[143] in R to fit elastic net regression models. Elastic net is a variation of linear regression. It attempts to solve for the coefficients of a linear equation that equates to the *line of best fit*. The line of best fit minimizes the squared distance between the line and each data point, which is also known as the least squares method. Ordinary linear regression is represented by the following equation:

$$\text{argmin} = \sum\left(y_a - y_p\right)^2$$

$$\text{argmin} = \sum(y_a - (B_1 x_1 + \cdots \beta_n x_n) - b)^2.$$

Argmin is the cost function in which we seek to minimize the answer with the given input arguments. Y_a is the value of the target label, Y_p is the prediction value calculated by the summation of the predictors *x* multiplied by β_n, a vector of coefficients which is determined by fitting the model, and *b* is the y-intercept[165]. The process of regularization involves the introduction of different variations of bias and penalties to find the solution to the equation above with the highest predictive accuracy. The alpha (α) argument controls the elastic net penalty. In *glmnet* α = 1 is lasso regression and α = 0 is ridge regression[143]. The ridge penalty shrinks the coefficients of correlated features toward each other while the lasso typically selects one of them and discards the others[31]. An α = 0.5 combines both the lasso and ridge for the elastic net, where correlated predictors may be selected or left out entirely as groups. The tuning parameter lambda (λ) controls the overall strength of the penalties. The best lambda is the value that minimizes the CV prediction error rate and it can be determined using the *cv.glmnet* function. The ridge penalty value (L2) adds a bias equal to the absolute value of the coefficients and the lasso penalty (L1) adds a bias equal to the squared value of the coefficients[31]. The general form of the elastic net equation is:

$$\text{argmin} = \sum(y_a - \beta x_n)^2 + \lambda_1 \sum|\beta| + \lambda_2 \sum \beta^2$$

We selected α = 0.5 (elastic net regression), *k*=10, and performed 10-fold CV using *cv.glmnet*. The mean squared error (MSE) was selected as the test error for the cv function and is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2,$$

where y_i is the predicted class label for the *i*th observation, $\hat{f}(x_i)$ is the prediction that $\hat{f}$ (our estimate for the unknown function) gives for the *i*th observation[148]. A small MSE indicates that the predicted responses are very close to the true responses and a large value occurs if there is a substantial difference between the predicted and true responses. The *k*-fold CV estimate is computed with the equation:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE.$$

After performing 10-fold CV we used *lambda.min*, the value of λ which gives the minimum mean cross-validated error[31], as the value for the lambda argument in *glmnet* to fit an elastic net model using the training data. The model was then evaluated on the test data to obtain a test MSE, which was converted to root-mean-square error (RMSE) by taking the square root of the MSE. The RMSE is used since it is measured in the same units as the response variable.

The 10-fold CV and testing process was repeated 1000 times for each feature selection subset to observe how the RMSE fluctuates between the trials. The seed was set to match the trial number for each iteration (e.g., trial 1 had a seed set to 1, trial 1000 had a seed set to 1000) and the test RMSE value for each trial is stored in a list for the relevant analysis. The first feature selection method involves restricting attention to the lowest variance probes in 25000 probe increments, which is equivalent to approximately 10% of the total probes. For example, the first subset of features includes all available probes, the second subset excludes the 25000 lowest variance probes, the third subset excludes the

lowest 50000 probes, and so on until the last subset which excludes the lowest 200000 probes. The second feature selection method restricts attention to the highest variance probes in 25000 probe increments. For example, the first subset of features includes all available features, the second subset excludes the 25000 highest variance probes, the third subset excludes the 50000 highest variance probes, and so on until the last subset which excludes the 200000 highest variance probes. The third feature selection method explores subsets of 100000 probes beginning at the lowest variance probe. The starting position of the subsequent subset increases in increments of 25000. For example, the first subset consists of the 100000 lowest variance probes, the second subset includes probes from index positions 25000 to 125000, the third subset includes probes from 50000 to 150000, and so on until the last set which includes probes from 150000 to 250000. The RMSE values were then charted in box plots for each method for visual summarization.

The best performing models were in the set that excluded the 150000 lowest variance probes. Using all the data (training and test), we used an elastic net model to regress a calibrated version of chronological age on 256801 CpG probes for a model with all features (all-features model) and on 106801 probes for a model which excluded the 150000 lowest variance probes (-150K model). Horvath's age transformation formula is a continuous, monotonically increasing function which can be inverted. It has a logarithmic dependence on age until adulthood (age 20), and it has a linear dependence on age after adulthood[23]. The calibration function is as follows:

F(age)=log(age+1)-log(adult.age+1) if age<= adult.age.

F(age)=(age-adult.age)/(adult.age+1) if age> adult.age.

The All-features linear regression model resultant from the elastic net regression includes coefficients b_0, b_1, ..., b_{252} and relates to transformed age as follows:

$$F(chronological\ age) = b_0 + b_1 CpG_1 + \cdots + b_{252} CpG_{252} + error.$$

The -150K linear regression model includes coefficients b_0, b_1, ..., b_{213} and relates to transformed age as follows:

$$F(chronological\ age) = b_0 + b_1 CpG_1 + \cdots + b_{213} CpG_{213} + error.$$

Based on the coefficient values from the All-features and -150K regression models, DNAmAge is estimated using the following respective equations:

$$DNAmAge = inverse.F(b_0 + b_1 CpG_1 + \cdots + b_{252} CpG_{252}),$$

$$DNAmAge = inverse.F(b_0 + b_1 CpG_1 + \cdots + b_{213} CpG_{213}).$$

The lambda values for the All-features and -150K elastic net regression were selected as 0.03005172 and 0.02613746, respectively, using 10-fold cross validation. The alpha values were set to 0.5 since we used an elastic net predictor. The final model for the All-features clock had 252 probes and the -150K clock had 213 probes. The two clocks were evaluated and compared using the datasets GSE101961 and GSE59157. The models were also applied to the PRAD and LIHC datasets to evaluate their performances on tissue types not trained on, and applied to KIRC, KIRP, and BRCA tumour samples to investigate applications to cancer datasets. The main workflow approach is outlined in figure 3.

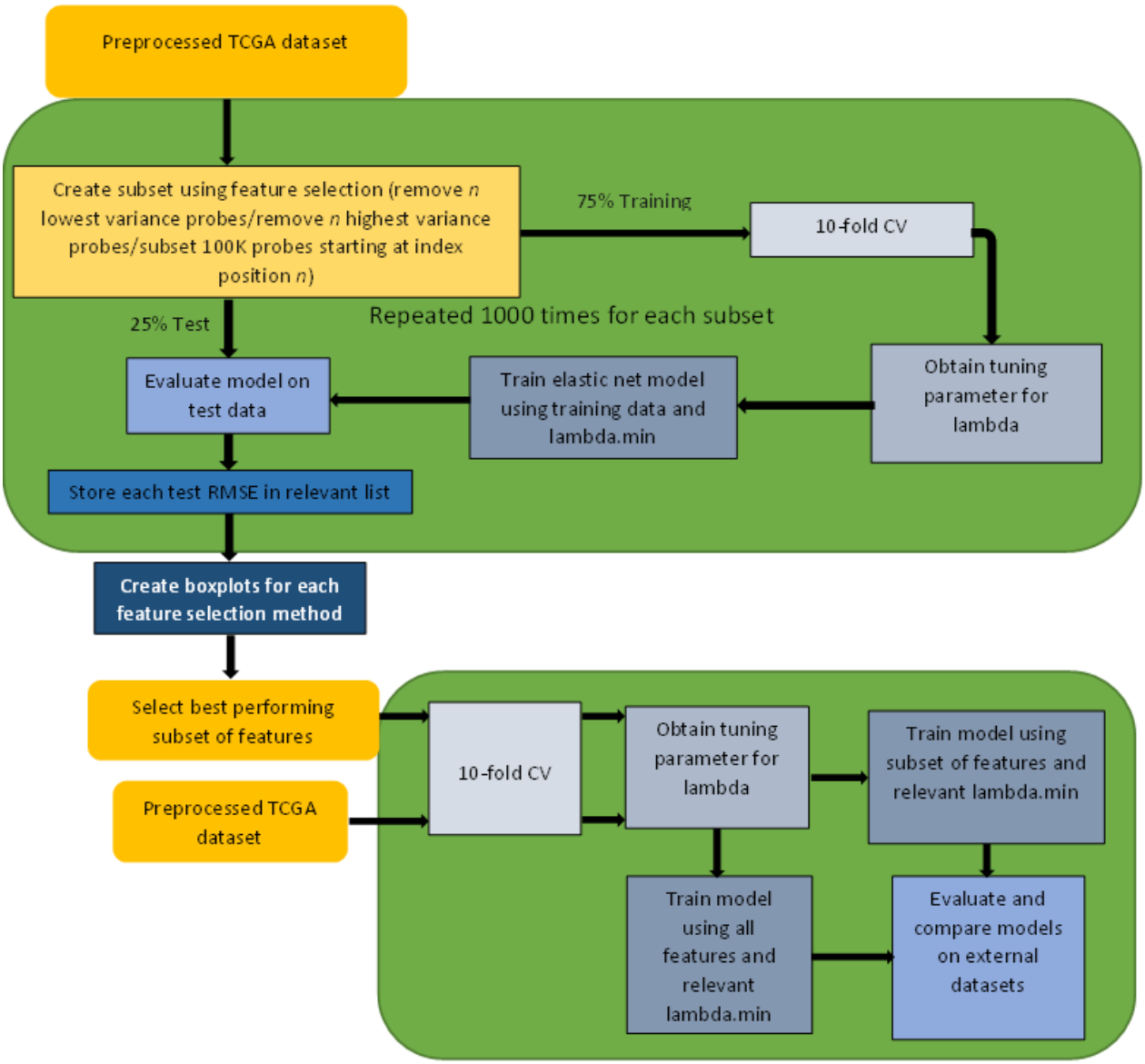

Figure 3. The general workflow for the model evaluation process. Feature selection was performed on our merged TCGA dataset and the resulting subset was split into a training and testing set for each iteration of 10-fold CV. The training data was used to choose the value of the tuning parameter lambda which gives the minimum mean cross-validated error, known as lambda.min. An elastic net regression model was fit using the training data and lambda.min and the model is evaluated on the test data. The test RMSE for each iteration is stored in a list and saved to a CSV file. This process is repeated 1000 times for each unique subset used in each feature selection method. The best feature selection subset was selected for 10-fold CV to obtain lambda.min. The final elastic net model was built using the best subset and its lambda.min. An elastic net model using all features, with no feature selection performed, was built as well. Both models were evaluated and compared on external datasets.

3 Results

To examine the effects of our feature selection methods, we applied our model evaluation process to a merged TCGA dataset of BRCA, KIRC, KIRP, LUAD, LUSC, and THCA (*n*=431). The age of the samples in the merged dataset range between 15 to 90 years old. Table 3 provides the median and the mean of the 1000 test RMSE values for the subsets in each feature selection approach, while Figure 4 provides a visual summarization in the form of box plots. Based on the mean and median RMSE values across all

subsets (Table 3), the best feature selection method is to exclude the lowest variance probes and the worst feature selection method is to exclude the highest variance probes. The lowest mean (5.790702) and median (5.789078) RMSE values are observed in the subset which excludes the 150000 lowest variance probes (Table 3A). The highest mean (9.143154) and median (9.103408) RMSE values are observed in the subset which excludes 200000 of the highest variance probes (Table 3B). From the 100K subset method (Table 3C), the mean and median RMSE values indicate that the best performing subset consists of the highest variance probes and the worst performing subset contains the lowest variance probes. Specifically, the subset containing probes from index position 1 to 100000 have a mean RMSE of 7.536545 and a median RMSE of 7.528993 across 1000 samples, and the subset containing probes from index position 150000 to 250000 have a mean RMSE of 5.799420 and a median RMSE of 5.796417. Figure 4A shows that there is a small improvement in RMSE by removing low variance probes up to 150000 for our dataset. There is a notable decrease in model performance when removing beyond the 25000 highest variance probes (Figure 4B). Figure 4C shows that test RMSE values improve with subsets of higher variance features.

We chose the overall best performing subset, which is the subset that excludes the 150000 lowest variance probes, to compare means of the test RMSE values with the All-features dataset. A quantile-quantile (QQ) plot shows the distribution of data values against the expected normal distribution. The data are normally distributed, as the points on each plot lie approximately on a straight line (Figure 5). A Welch two sample t-test indicates that the means of the two subset samples differ (p=1.855e-05) and therefore, there is a significant difference between the two groups.

A	**Exclude *n* lowest variance probes (10^3)**								
	0	25	50	75	100	125	150	175	200
Mean RMSE	5.880109	5.884252	5.884519	5.883096	5.869347	5.810648	5.790702	5.867459	6.043872
Median RMSE	5.876341	5.860561	5.870075	5.866863	5.855069	5.795560	5.789078	5.868769	6.047802
B	**Exclude *n* highest variance probes (10^3)**								
	25	50	75	100	125	150	175	200	
Mean RMSE	5.876454	6.093485	6.192969	6.334583	6.457064	7.327633	7.859929	9.143154	
Median RMSE	5.864047	6.079400	6.169910	6.314939	6.427752	7.318194	7.827536	9.103408	
C	**100K subset range by lowest index position (K = thousand)**								
	1	25K	50K	75K	100K	125K	150K		
Mean RMSE	7.536545	6.763968	6.436773	6.235975	6.163197	5.912452	5.799420		
Median RMSE	7.528993	6.745027	6.421636	6.219771	6.160157	5.889370	5.796417		

Table 3. Mean and median RMSE values based on 1000 trials for the subsets in each feature selection approach. A) Feature selection approach which excludes lowest variance probes to create subsets. Mean and median values are marginally different across all subsets. The subset excluding the 150000 lowest variance probes is associated with the lowest mean and median RMSE values. B) Feature selection approach which excludes the highest variance probes to create subsets. The mean and median RMSE values increase as more high variance probes are excluded. The best subset in this approach excludes 25000 high variance probes, which is the least amount of high variance probes excluded. C) Feature selection approach which creates subsets of 100000 probes. Lower variance probes have a lower index position. The mean and median RMSE values decrease as the subsets include higher variance probes. The best subset in this approach ranges from index position 150K to 250K.

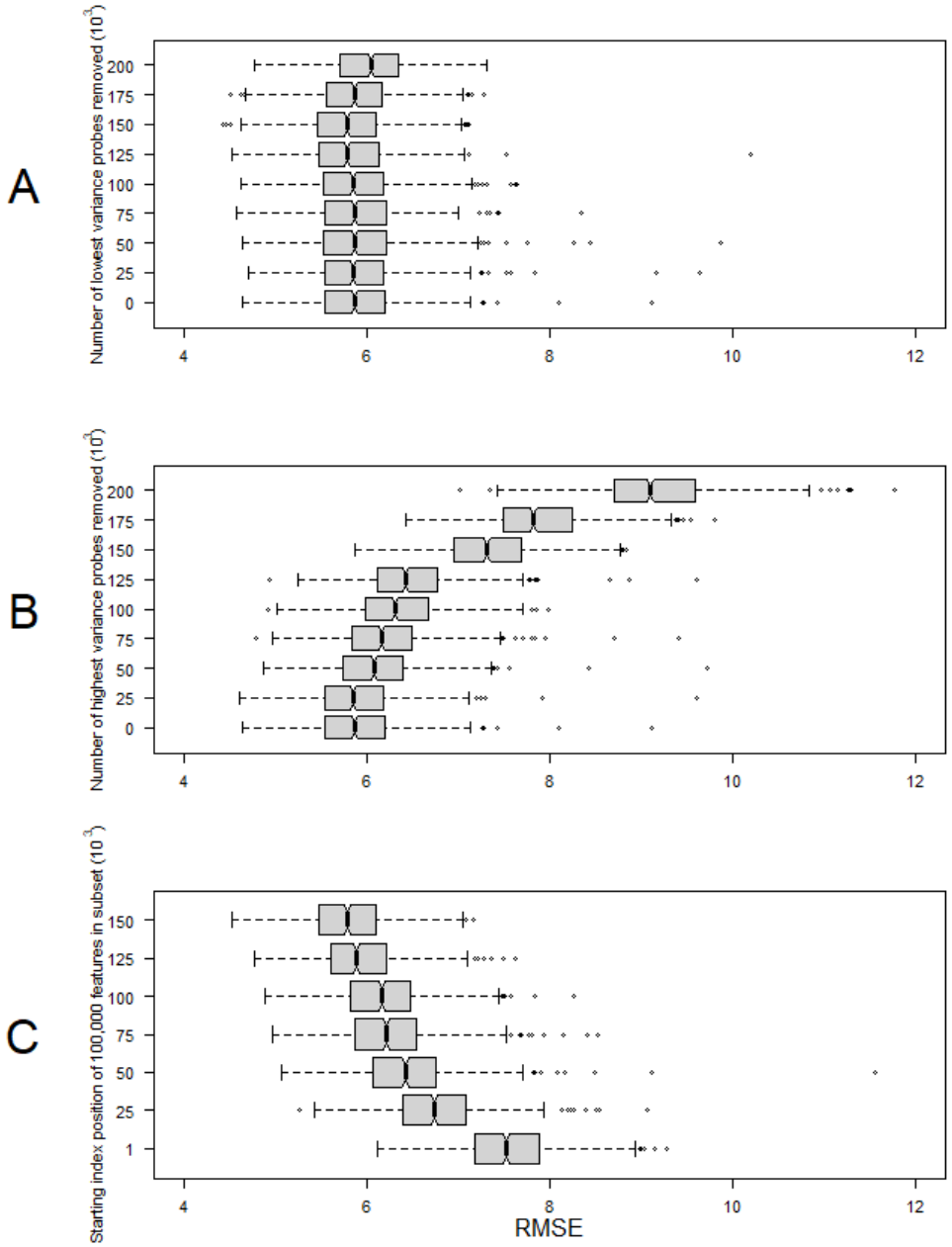

Figure 4. Box plots of the test RMSE values from the 1000 trials of 10-fold CV and model fitting using the corresponding lambda.min for each subset. A) Box plots for the feature selection method which excludes low variance probes from subsets. The lowest median is found in the subset which excludes the 150000 lowest variance probes, while the highest median is found in the subset which excludes the 200000 lowest variance probes. The interquartile ranges are relatively similar, as shown by the length of the boxes. The overall range of values is greater in the subsets with less than 150000 low variance probes excluded when considering the outliers. There is no significant skewness observed in any of the sets of data. B) Box plots for the feature selection method which excludes high variance probes from subsets. The lowest median is observed in the subset which excludes 25000 of the highest variance probes. The median value increases in subsets with a greater amount of high variance probes excluded. The interquartile ranges are reasonably similar for most of the subsets but appear to slightly increase above 125000 and is most noticeable in the 200000-probe subset. There is no significant skewness observed in any of these datasets. C) Box plots for the feature selection method which creates subsets of 100000 probes. The median of the subsets ranging from index position 150000 to 250000 (150K to 250K) contain the highest variance probes out of all 100000-probe subsets compared. The interquartile ranges are similar, but the 150K to 250K subset has the lowest range of test RMSE values when considering outliers. There is no significant skewness observed in any of the datasets. Overall, the median value is lower in subsets containing higher variance probes.

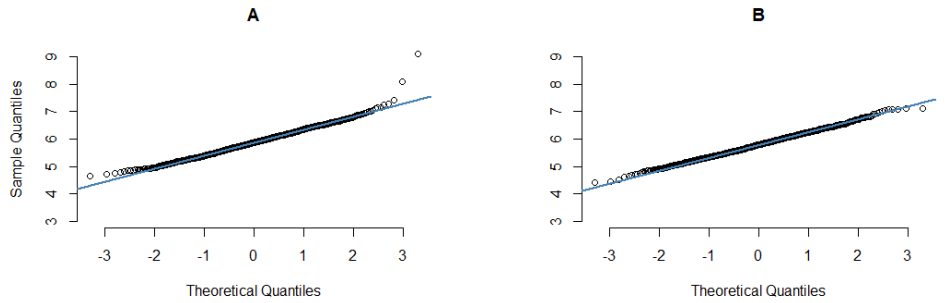

Figure 5. Quantile-quantile (QQ) plots, or normal probability plots, which show the distribution of the data against expected normal distribution. Normally distributed data should lie approximately on the straight blue lines. A) Normal QQ plot for the test RMSE values of the all features dataset. The data points lie on a straight line, except for several outliers. The data are normally distributed. B) Normal QQ plot for the test RMSE values of the subset which excludes the 150000 lowest variance probes. The data points lie on a straight line and are normally distributed.

The All-features model represents a model without feature selection applied. This model uses all the 256801 features to train an elastic net model. The clock which results from this approach is based on 252 CpGs. The 252 clock CpGs for the All-features model can be divided according to their correlation with age into two sets. The 132 positively and 120 negatively correlated CpGs become hypermethylated and hypomethylated with age, respectively. The -150K model uses the feature selection method of excluding the 150000 lowest variance probes, which results in a model trained on 106801 features. The resulting clock from this approach is based on 213 CpGs, of which 112 and 101 are positively and negatively correlated with age, respectively. Figure 6 is a combined histogram and density plot of the clock CpGs ordered according to their standard deviation positions. The All-features clock has 155 CpGs with standard deviation positions below 150000 and 97 probes from 150000 to 256801, with 81 of the 97 CpGs common to both the All-Features and the -150K model. The -150K model has 152 CpGs with standard deviation positions between 150000 and 200000, which is approximately 71% of its CpGs.

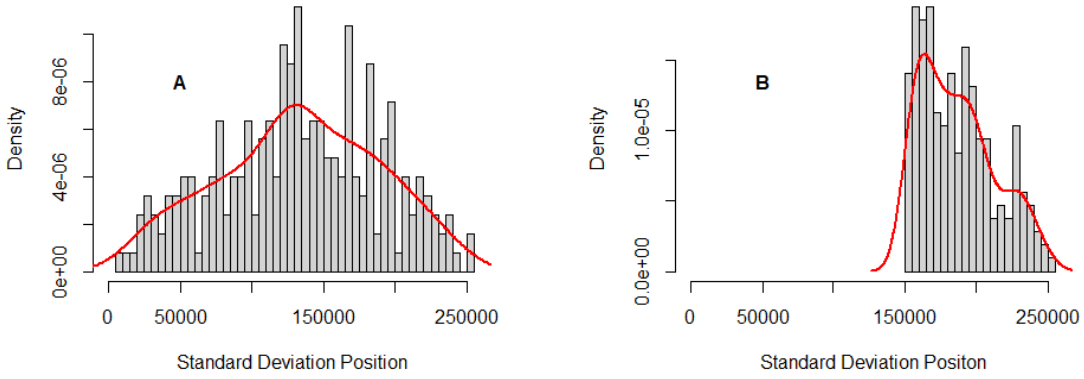

Figure 6. Combined histogram and density plot of standard deviation positions for clock CpGs. A) The All-features model is based on 252 clock CpGs, of which 155 are under the 150000 position and 97 are at or above this position. B) The -150K model is based on 213 clock CpGs, of which 152 have standard deviation positions between 150000 and 200000.

We evaluated and compared the All-features model and the -150K model on several datasets based on the Illumina 450K platform. GSE101961 is a methylation profiling of breast tissue taken from 121 cancer-free women aged 17 to 76 years[152]. GSE59157 is a methylation profiling of normal kidney (*n*=36), nephrogenic rest (*n*=22), and Wilms tumour (*n*=37) taken from children aged 10 to 144 months[166]. TCGA

cancer datasets used for our evaluation include KIRC (*n*=319), KIRP (*n*=217), and BRCA (*n*=756). Ages for the TCGA cancer datasets range from 26 to 29 years, 28 to 88 years, and 26 to 29 years, respectively. We also compared the performance of the two models on TCGA PRAD (*n*=50) and LIHC (*n*=50), which are two tissue types they were not trained on. The ages of the PRAD samples range from 44 to 72 years and for LIHC they range from 20 to 81 years.

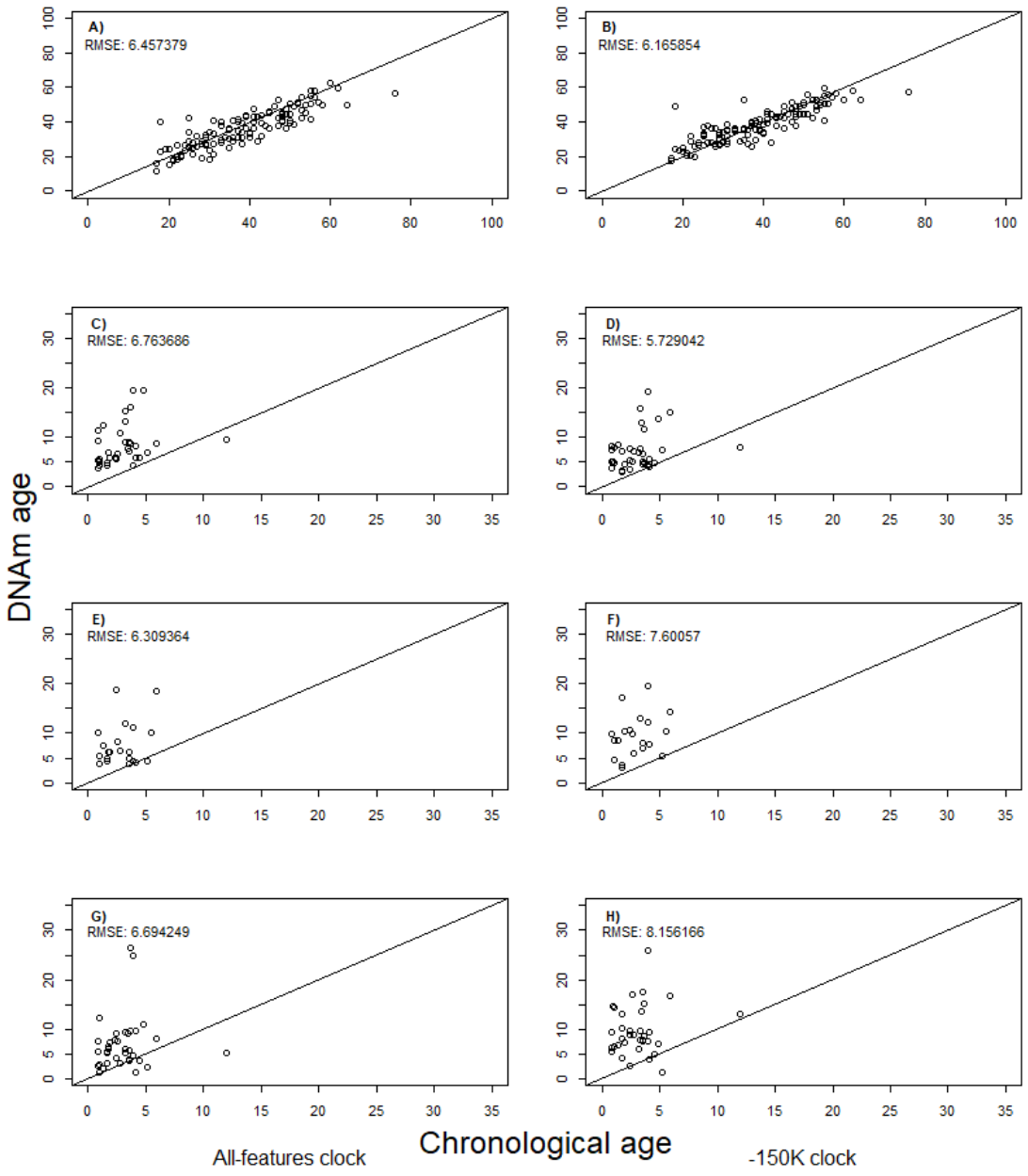

Figure 7. Scatterplots of normal samples' chronological age vs DNAm age for the All-features model (A, C, E, G) and -150K model (B, D, F, H) on the GEO datasets GSE101961 (A-B) and GSE59157 (C-H). The line in each plot represents the trivial result if the correspondence was exact. (A-B) GSE101961 normal breast samples (n=121). (C-D) GSE59157 normal kidney samples (n=36). (E-F) GSE59157 nephrogenic rest kidney samples (n=22). (G-H) GSE59157 Wilms tumour kidney samples (n=37).

For the GSE101961 dataset, the All-features clock achieved a RMSE of 6.457379 and the -150 clock achieved a lower RMSE of 6.165854 (Figure 7A-B). For the GSE59157 dataset, the -150K clock had a lower RMSE for the normal samples, but a higher RMSE for nephrogenic rest and Wilms tumour samples than the All-features clock. The All-features clock achieved scores of 6.763686 for the normal samples, 6.309364 for the nephrogenic rest samples, and 6.694249 for the Wilms tumour samples (Figure 7C, E, G). The -150K clock achieved a lower score than the All-features clock for the normal samples with a RMSE of 5.729042, but a higher score for both the nephrogenic rest samples and the Wilms tumour samples with a RMSE 7.60057 and 8.156166, respectively (Figure 7D, F, H).

Figure 8 shows the scatterplots of normal samples' chronological age vs DNAm age of the two models for the datasets TCGA PRAD and LIHC. Overall, both models did not perform very well on normal prostate and liver samples. The All-features model achieved a RMSE of 11.26036 for the PRAD samples (Figure 8A) and 17.47927 for the LIHC samples (Figure 8C). The -150K model had RMSE values of 11.8528 (Figure 8B) and 10.89658 (Figure 8D) for the PRAD and LIHC samples, respectively.

Both models achieved high RMSE values for TCGA cancer datasets (Figure 9), which suggest epigenetic age acceleration. The All-features clock had RMSE values of 26.18829 for the KIRC tumour samples (Figure 9A), 18.98291 for the KIRP tumour samples (Figure 9C), and 30.12031 for the BRCA tumour samples (Figure 9E). The -150K clock achieved RMSE values of 22.24626 for the KIRC tumour samples (Figure 9B), 18.43928 for the KIRP tumour samples (Figure 9D), and 22.10403 for the BRCA tumour samples (Figure 9F). The difference between DNAm age and chronological age, or age acceleration, for each set of tumour samples is shown in Figure 10. For the KIRC tumour samples, there is generally positive age acceleration (Figure 10A-B). The KIRP tumour samples show similar amounts of positive and negative age acceleration (Figure 10C-D), while the BRCA tumour samples appear to show more negative age acceleration (Figure 10E-F). Additionally, the -150K clock had 210 of 756 BRCA tumour predictions within its test RMSE (6.165854) from the GSE101961 dataset, while the All-features clock had 163 of 756 DNAm predictions within its test RMSE (6.4573979). Overall, the All-features clock detects greater age acceleration in tumour samples than the -150K clock.

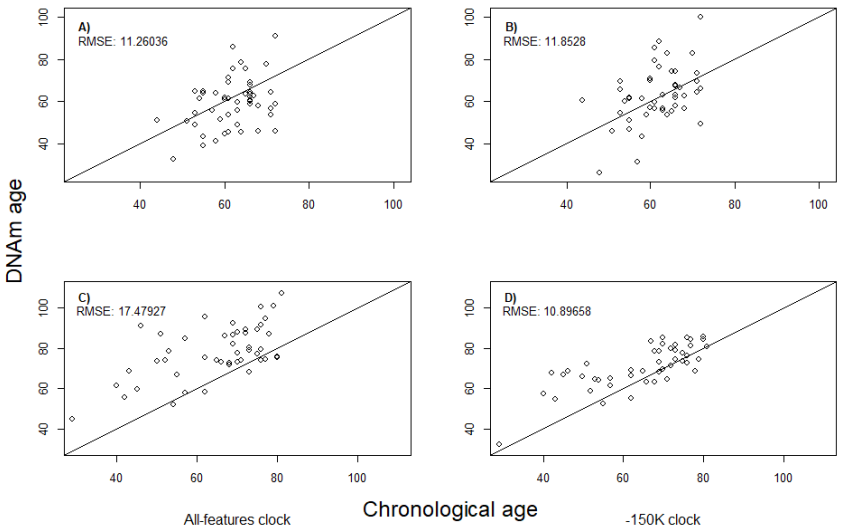

Figure 8. Scatterplots of normal samples' chronological age vs DNAm age for the All-features model (A, C) and -150K model (B, D) for TCGA PRAD (A-B) and LIHC (C-D). The line in each plot represents the trivial result if the correspondence was exact. (A-B) PRAD normal samples (n=50). (C-D) LIHC normal samples (n=50).

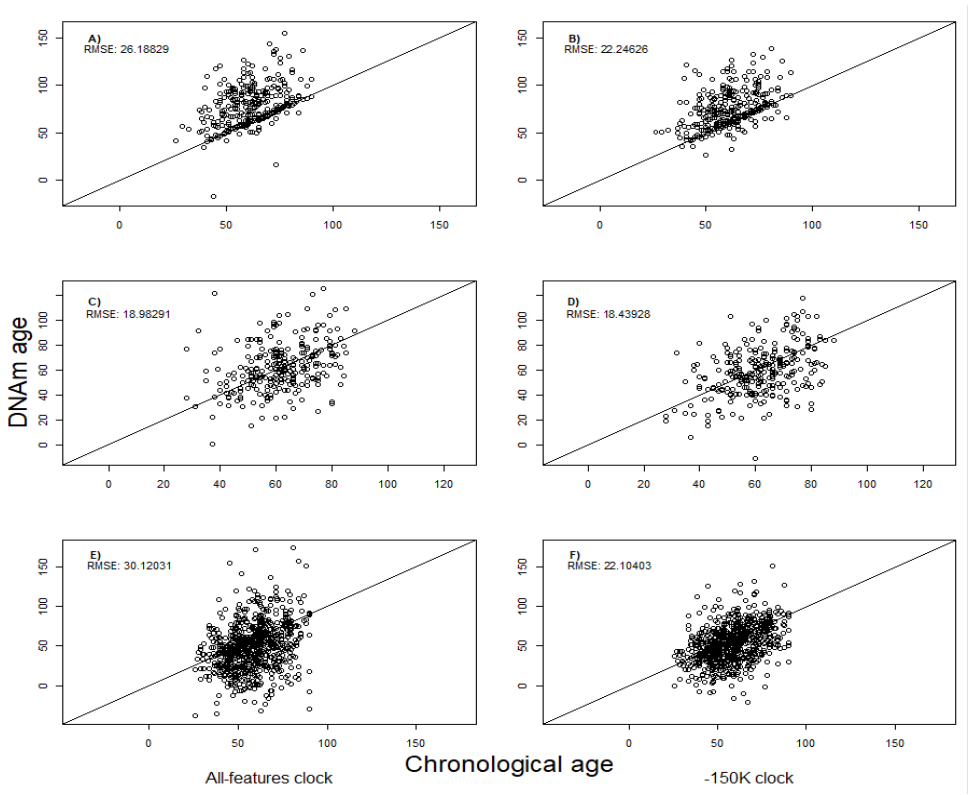

Figure 9. Scatterplots of tumour samples' chronological age vs DNAm age for the All-features model (A, C, E) and -150K model (B, D, F) for TCGA KIRC (A-B), KIRP (C-D), and BRCA (E-F). The line in each plot represents the trivial result if the correspondence was exact. (A-B) KIRC tumour samples (n=319). (C-D) KIRP tumour samples (n=271). (E-F) BRCA tumour samples (n=756).

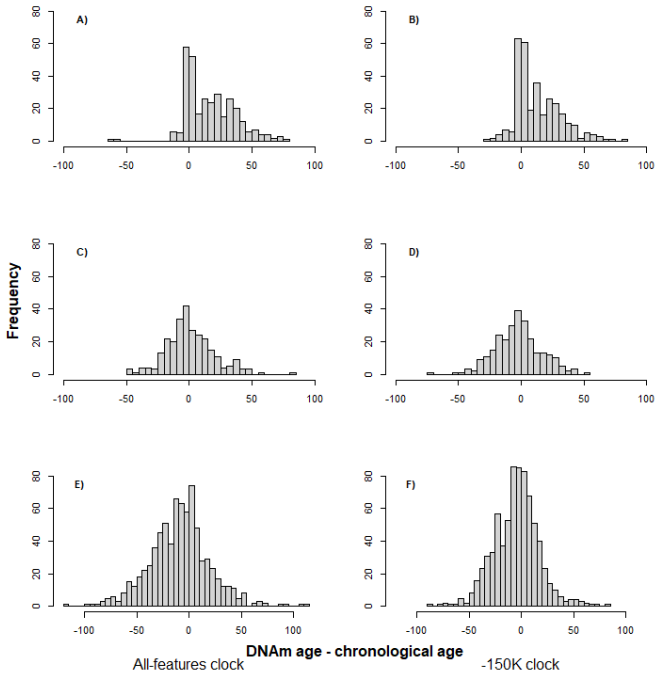

Figure 10. Histograms of difference between DNAm age and chronological age (DNAm age - chronological age) of tumour samples for the All-features clock (A, C, E) and the -150K clock (B, D, F). (A-B) KIRC tumour samples (n=319). (C-D) KIRP tumour samples (n=217). (E-F) BRCA tumour samples (n=756).

4 Discussion

Overall, we find that feature selection prior to elastic net regression modeling can produce a more accurate clock with fewer CpGs for predicting the DNAm age of normal samples. This is in comparison to the common approach of creating a clock by using all measured CpG sites in elastic net regression modeling, such as for our All-features model. When we restricted attention to low variance CpGs, we found a small, but observable improvement in our cross-validated fits between 125000 and 150000 (Table 3A, Figure 4A). Any less than this threshold resulted in negligible improvement, while excluding any more than 200000 of the lowest variance CpGs appeared to yield poorer predictive results. Although we describe it as removing low variance features for this approach, we must consider the relatively high variance of CpG beta-values near the 200000-index position.

Our internal validation process of all three feature selection approaches helped us to identify a common theme regarding the variance of our features. Removing low variance CpGs had a marginal improvement on model performance only within a certain threshold or range, as 25previously described. Removing high variance CpGs decreased the predictive accuracy of models (Table 3B, Figure 4B). As more of the highest variance probes were removed, we saw a decrease in model performances as indicated by higher RMSE values. However, excluding the 25000 highest variance probes did not appear to decrease model performance. Using subsets of 100000 features, we found that the best performing models included higher variance CpGs (Table 3C, Figure 4C). The best performing models in this approach were built from the subset containing the 150000 to 250000 highest variance CpGs out of the total 256801. The mean (5.796417) and median (5.799420) RMSE values of this subset were comparable to the -150K subset's median (5.790702) and mean (5.789078). Together, these results suggest that the lower variance CpGs do not meaningfully contribute to the predictive capability of models for estimating DNAm age in normal samples, while higher variance CpGs are strongly correlated with chronological age.

The -150K clock performed better than the All-features clock on the 121 normal breast tissue samples of the external dataset GSE101961 (Figure 7A-B). This was achieved with the -150K clock using 39 less CpGs than the All-features clock, which is approximately a 15% difference. Without any feature selection the elastic net regression selected 252 CpGs from the 256801 total CpGs, 81 of which overlap with the -150K clock's 213 CpGs. These results suggest that the feature selection method of excluding low variance features reliably selects good CpGs to use in the construction of clocks for estimating the DNAm of normal samples. This feature selection approach produces a simpler and more interpretable clock in comparison to solely relying on the elastic net for both feature selection and prediction. We also applied our models to datasets of different sample types. The DNAm predictions for both clocks on the PRAD and LIHC datasets resulted in poor predictions and high RMSE values (Figure 8). These clocks are not reliable for DNAm age prediction on tissue types other than those trained on, as the identified CpGs are likely to be dataset or tissue specific.

In addition to applying our clocks to normal samples, we also investigated the effects of our feature selection approach on cancer samples by testing the clocks on TCGA KIRC, KIRP, and BRCA tumour tissue. Across the three cancer datasets, the -150K clock achieved lower RMSE values than the All-features clock (Figure 9). However, the RMSE values for both models on the cancer datasets were much higher in general than any of the other datasets we tested. As previously mentioned, epigenetic age acceleration occurs when the predicted DNAm age is greater, or older than, the corresponding

chronological age. Cancer tissues have been shown to exhibit both positive and negative age acceleration. For example, luminal breast cancer tissue exhibits strong positive age acceleration, which contrasts with the negative age acceleration observed in basal breast cancers[167]. There is a weak correlation between DNAm age and chronological age, and a weak correlation between age acceleration and tumour morphology (grade and stage)[23]. For the KIRC tumour samples, our results show that there is typically positive age acceleration and more age acceleration overall for the All-features model (Figure 10A-B). For the KIRP tumour samples, our results show both positive and negative age acceleration in approximately similar amounts (Figure 10C-D). Results for the BRCA tumour samples had the most notable difference between the two clocks. The All-features clock demonstrated a greater amount of age acceleration than the -150K clock, as shown by the frequencies in Figure 10E-F. We also see age acceleration in our results for the GSE59157 dataset, where the nephrogenic rest (Figure 7C-D) and Wilms tumour (Figure7E-F) samples have higher RMSE values than the normal samples (Figure 7A-B). However, we interpret these results with caution as the sample sizes were relatively small and since our training data age range (15 to 90 years) did not include the ages in the GSE59157 dataset (0 to 12 years).

While a lower RMSE score is indicative of a more accurate epigenetic clock for predicting DNAm age, this is not desired when the clock is applied to cancerous tissue. Age acceleration is anticipated in cancer tissue[23]. If a clock is unable to detect an adequate amount of age acceleration for cancerous tissue, then a sample may be given a false negative result. Age acceleration in cancer tissue should be greater than in normal samples. The -150K clock has 210 of 756 DNAm predictions for the BRCA tumour samples within its RMSE (6.165854) based on the normal breast tissue samples of GSE101961, while the All-features clock has 163 of 756 DNAm predictions within its RMSE (6.4573979) based on the same datasets, respectively. Hence, more age acceleration is demonstrated in the All-features clock than the -150K clock for the BRCA cancer samples. A possible explanation for this difference between the clocks is that the threshold used for the -150K feature selection method excludes CpGs that are associated with specific mutations, which in turn are associated with accelerated DNAm age in breast cancer. Horvath's original clock publication found that somatic mutations in steroid receptors, such as estrogen receptors or progesterone receptors, are associated with accelerated DNAm age in breast cancer[23]. Additionally, the variance of a feature may ignore the relationship between the feature and response. While a CpG may have a low variance, the relationship between the CpG and chronological age may be powerful within that range.

The feature selection methods we investigated here gives insight into the potential improvements that could be made to existing clocks built exclusively using elastic net regression modeling. Typically, with the elastic net we see more CpGs selected with a greater amount of training (Table 1). Various feature selection methods have been previously shown to outperform stock selection methods, yielding clocks with a high accuracy for age prediction while utilizing a low number of CpG sites[165]. We demonstrate that a clock based on fewer CpGs and with improved accuracy can be achieved by simply removing lower variance features before elastic net regression. This approach can be applied to any clock based on the elastic net. However, we note that utilizing this feature selection method may result in a lesser capability for predicting age acceleration in cancer samples. Furthermore, determination of the variance threshold requires a trial-and-error approach relative to each training dataset.

Our study was limited by the number of normal samples available to test our models on and by the time allotted for our research. An exhaustive search within the scope of the project yielded only the GSE101961 dataset and GSE59157 for normal sample testing, of which the latter we take caution in

interpreting the results. Many external datasets were missing the necessary data for use in our clocks such as probe measurements and patient ages. Given more time, we would investigate if we could build a clock with comparable accuracy using the 81 overlapping CpG sites between the All-features clock and -150K clock. Ideally, we would also investigate the chromosome locations of the different CpGs between the two clocks to confirm if any breast cancer or tumour associated CpG sites were excluded in the -150K clock threshold.

We recommend that the feature selection methods investigated here be applied to models trained on different datasets, such as Horvath's dataset or any dataset which is associated with an elastic net model, to examine how removing low variance features affects clocks trained on different sample sizes and tissue types. We also suggest that models using this feature selection approach should evaluate clock performances on tissue or cell types known to exhibit age acceleration, such as cancer tissue, since key CpG sites associated with age acceleration could be excluded.

5 References

1. Kane AE, Sinclair DA. Epigenetic changes during aging and their reprogramming potential. *Crit Rev Biochem Mol Biol*. 2019;54(1):61-83. doi:10.1080/10409238.2019.1570075

2. Pal S, Tyler JK. Epigenetics and aging. *Sci Adv*. 2016;2(7):e1600584. doi:10.1126/sciadv.1600584

3. Hardy TM, Tollefsbol TO. Epigenetic diet: impact on the epigenome and cancer. *Epigenomics*. 2011;3(4):503-518. doi:10.2217/epi.11.71

4. Tiffon C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. *Int J Mol Sci*. 2018;19(11):3425. doi:10.3390/ijms19113425

5. Santos JM, Tewari S, Benite-Ribeiro SA. The effect of exercise on epigenetic modifications of PGC1: The impact on type 2 diabetes. *Med Hypotheses*. 2014;82(6):748-753. doi:10.1016/j.mehy.2014.03.018

6. Fernandes J, Arida RM, Gomez-Pinilla F. Physical Exercise as an Epigenetic Modulator of Brain Plasticity and Cognition. *Neurosci Biobehav Rev*. 2017;80:443-456. doi:10.1016/j.neubiorev.2017.06.012

7. Denham J, Marques FZ, O'Brien BJ, Charchar FJ. Exercise: Putting Action into Our Epigenome. *Sports Med*. 2014;44(2):189-209. doi:10.1007/s40279-013-0114-1

8. Alegría-Torres JA, Baccarelli A, Bollati V. Epigenetics and lifestyle. *Epigenomics*. 2011;3(3):267-277. doi:10.2217/epi.11.22

9. Reamon-Buettner SM, Mutschler V, Borlak J. The next innovation cycle in toxicogenomics: Environmental epigenetics. *Mutat Res Mutat Res*. 2008;659(1):158-165. doi:10.1016/j.mrrev.2008.01.003

10. Gibney ER, Nolan CM. Epigenetics and gene expression. *Heredity*. 2010;105(1):4-13. doi:10.1038/hdy.2010.54

11. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295. doi:10.1016/j.ygeno.2011.07.007

12. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology*. 2013;38(1):23-38. doi:10.1038/npp.2012.112

13. Cheung P, Lau P. Epigenetic Regulation by Histone Methylation and Histone Variants. *Mol Endocrinol*. 2005;19(3):563-573. doi:10.1210/me.2004-0496

14. Peschansky VJ, Wahlestedt C. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics*. 2014;9(1):3-12. doi:10.4161/epi.27473

15. Dupont C, Armant DR, Brenner CA. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Semin Reprod Med*. 2009;27(5):351-357. doi:10.1055/s-0029-1237423

16. Frías-Lasserre D, Villagra CA. The Importance of ncRNAs as Epigenetic Mechanisms in Phenotypic Variation and Organic Evolution. *Front Microbiol*. 2017;8:2483. doi:10.3389/fmicb.2017.02483

17. Bibikova M, Le J, Barnes B, et al. Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics*. 2009;1(1):177-200. doi:10.2217/epi.09.14

18. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The Hallmarks of Aging. *Cell*. 2013;153(6):1194-1217. doi:10.1016/j.cell.2013.05.039

19. Bocklandt S, Lin W, Sehl ME, et al. Epigenetic Predictor of Age. *PLoS ONE*. 2011;6(6):e14821. doi:10.1371/journal.pone.0014821

20. Rakyan VK, Down TA, Maslau S, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res*. 2010;20(4):434-439. doi:10.1101/gr.103101.109

21. Horvath S, Zhang Y, Langfelder P, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol*. 2012;13(10):R97. doi:10.1186/gb-2012-13-10-r97

22. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*. 2005;102(30):10604-10609. doi:10.1073/pnas.0500398102

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

24. Hannum G, Guinney J, Zhao L, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell*. 2013;49(2):359-367. doi:10.1016/j.molcel.2012.10.016

25. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*. 2018;10(4):573-591. doi:10.18632/aging.101414

26. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*. 2019;11(2):303-327. doi:10.18632/aging.101684

27. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17:208. doi:10.1186/s13059-016-1066-1

28. Zhang Q, Vallerga CL, Walker RM, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med*. 2019;11:54. doi:10.1186/s13073-019-0667-1

29. Carless MA. Determination of DNA Methylation Levels Using Illumina HumanMethylation450 BeadChips. In: Chellappan SP, ed. *Chromatin Protocols*. Methods in Molecular Biology. Springer; 2015:143-192. doi:10.1007/978-1-4939-2474-5_10

30. Illumina. Illumina Methylation BeadChips Achieve Breadth of Coverage Using 2 Infinium® Chemistries. Published online 2015.

https://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf

31. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

32. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet*. 2018;19(6):371-384. doi:10.1038/s41576-018-0004-3

33. Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. *Aging*. 2011;3(10):1018-1027.

34. Koch CM, Joussen S, Schellenberg A, Lin Q, Zenke M, Wagner W. Monitoring of cellular senescence by DNA-methylation at specific CpG sites. *Aging Cell*. 2012;11(2):366-369. doi:10.1111/j.1474-9726.2011.00784.x

35. Simpson DJ, Chandra T. Epigenetic age prediction. *Aging Cell*. 2021;20(9):e13452. doi:10.1111/acel.13452

36. Bergsma T, Rogaeva E. DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. *Neurosci Insights*. 2020;15:2633105520942221. doi:10.1177/2633105520942221

37. DNA methylation-based age clocks: From age prediction to age reversion. *Ageing Res Rev*. 2021;68:101314. doi:10.1016/j.arr.2021.101314

38. Weidner CI, Lin Q, Koch CM, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol*. 2014;15(2):R24. doi:10.1186/gb-2014-15-2-r24

39. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum Mol Genet*. 2014;23(5):1186-1201. doi:10.1093/hmg/ddt531

40. Polanowski AM, Robbins J, Chandler D, Jarman SN. Epigenetic estimation of age in humpback whales. *Mol Ecol Resour*. 2014;14(5):976-987. doi:10.1111/1755-0998.12247

41. Huang Y, Yan J, Hou J, Fu X, Li L, Hou Y. Developing a DNA methylation assay for human age prediction in blood and bloodstain. *Forensic Sci Int Genet*. 2015;17:129-136. doi:10.1016/j.fsigen.2015.05.007

42. Zbieć-Piekarska R, Spólnicka M, Kupiec T, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci Int Genet*. 2015;17:173-179. doi:10.1016/j.fsigen.2015.05.001

43. Yang Z, Wong A, Kuh D, et al. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol*. 2016;17:205. doi:10.1186/s13059-016-1064-3

44. Knight AK, Craig JM, Theda C, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol*. 2016;17:206. doi:10.1186/s13059-016-1068-z

45. Golbabapour S, Majid NA, Hassandarvish P, Hajrezaie M, Abdulla MA, Hadi AHA. Gene Silencing and Polycomb Group Proteins: An Overview of their Structure, Mechanisms and Phylogenetics. *OMICS J Integr Biol*. 2013;17(6):283-296. doi:10.1089/omi.2012.0105

46. Zhang Y, Wilson R, Heiss J, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat Commun*. 2017;8(1):14617. doi:10.1038/ncomms14617

47. Cho S, Jung SE, Hong SR, et al. Independent validation of DNA-based approaches for age prediction in blood. *Forensic Sci Int Genet*. 2017;29:250-256. doi:10.1016/j.fsigen.2017.04.020

48. Wang T, Tsui B, Kreisberg JF, et al. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol*. 2017;18:57. doi:10.1186/s13059-017-1186-2

49. Petkovich DA, Podolskiy DI, Lobanov AV, Lee SG, Miller RA, Gladyshev VN. Using DNA methylation profiling to evaluate biological age and longevity interventions. *Cell Metab*. 2017;25(4):954-960.e6. doi:10.1016/j.cmet.2017.03.016

50. Stubbs TM, Bonder MJ, Stark AK, et al. Multi-tissue DNA methylation age predictor in mouse. *Genome Biol*. 2017;18:68. doi:10.1186/s13059-017-1203-5

51. Thompson MJ, vonHoldt B, Horvath S, Pellegrini M. An epigenetic aging clock for dogs and wolves. *Aging*. 2017;9(3):1055-1068. doi:10.18632/aging.101211

52. Ito H, Udono T, Hirata S, Inoue-Murayama M. Estimation of chimpanzee age based on DNA methylation. *Sci Rep*. 2018;8:9998. doi:10.1038/s41598-018-28318-9

53. Meer MV, Podolskiy DI, Tyshkovskiy A, Gladyshev VN. A whole lifespan mouse multi-tissue DNA methylation clock. Kaeberlein M, Tyler JK, Wagner W, Adams P, eds. *eLife*. 2018;7:e40675. doi:10.7554/eLife.40675

54. Thompson MJ, Chwiałkowska K, Rubbi L, et al. A multi-tissue full lifespan epigenetic clock for mice. *Aging*. 2018;10(10):2832-2854. doi:10.18632/aging.101590

55. Horvath S, Oshima J, Martin GM, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging*. 2018;10(7):1758-1775. doi:10.18632/aging.101508

56. Youn A, Wang S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. *Epigenetics*. 2018;13(2):192-206. doi:10.1080/15592294.2017.1389361

57. Lu AT, Seeboth A, Tsai PC, et al. DNA methylation-based estimator of telomere length. *Aging*. 2019;11(16):5895-5923. doi:10.18632/aging.102173

58. McEwen LM, O'Donnell KJ, McGill MG, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc Natl Acad Sci U S A*. 2020;117(38):23329-23335. doi:10.1073/pnas.1820843116

59. Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci Int Genet*. 2019;38:1-8. doi:10.1016/j.fsigen.2018.09.010

60. Montpetit AJ, Alhareeri AA, Montpetit M, et al. Telomere Length: A Review of Methods for Measurement. *Nurs Res*. 2014;63(4):289-299. doi:10.1097/NNR.0000000000000037

61. Little TJ, O'Toole AN, Rambaut A, Chandra T, Marioni R, Pedersen AB. Methylation-Based Age Estimation in a Wild Mouse. Published online July 16, 2020:2020.07.16.203687. doi:10.1101/2020.07.16.203687

62. Wang T, Ma J, Hogan AN, et al. Quantitative Translation of Dog-to-Human Aging by Conserved Remodeling of the DNA Methylome. *Cell Syst*. 2020;11(2):176-185.e6. doi:10.1016/j.cels.2020.06.006

63. Guevara EE, Lawler RR, Staes N, et al. Age-associated epigenetic change in chimpanzees and humans. *Philos Trans R Soc B Biol Sci*. 2020;375(1811):20190616. doi:10.1098/rstb.2019.0616

64. Lowe R, Danson AF, Rakyan VK, et al. DNA methylation clocks as a predictor for ageing and age estimation in naked mole-rats, Heterocephalus glaber. *Aging*. 2020;12(5):4394-4406. doi:10.18632/aging.102892

65. Anastasiadi D, Piferrer F. A clockwork fish: Age prediction using DNA methylation-based biomarkers in the European seabass. *Mol Ecol Resour*. 2020;20(2):387-397. doi:10.1111/1755-0998.13111

66. Mayne B, Korbie D, Kenchington L, Ezzy B, Berry O, Jarman S. A DNA methylation age predictor for zebrafish. *Aging*. 2020;12(24):24817-24835. doi:10.18632/aging.202400

67. Levine M, McDevitt RA, Meer M, et al. A rat epigenetic clock recapitulates phenotypic aging and co-localizes with heterochromatin. Suh Y, Tyler JK, Laird P, Schumacher B, eds. *eLife*. 2020;9:e59201. doi:10.7554/eLife.59201

68. Horvath S, Singh K, Raj K, et al. Reversing age: dual species measurement of epigenetic age with a single clock. Published online May 8, 2020:2020.05.07.082917. doi:10.1101/2020.05.07.082917

69. Sailer LL, Haghani A, Zoller JA, Li CZ, Ophir AG, Horvath S. Pair bonding slows epigenetic aging and alters methylation in brains of prairie voles. Published online September 26, 2020:2020.09.25.313775. doi:10.1101/2020.09.25.313775

70. Lemaître JF, Rey B, Gaillard JM, et al. Epigenetic clock and DNA methylation studies of roe deer in the wild. Published online September 22, 2020:2020.09.21.306613. doi:10.1101/2020.09.21.306613

71. Correia Dias H, Cordeiro C, Corte Real F, Cunha E, Manco L. Age Estimation Based on DNA Methylation Using Blood Samples From Deceased Individuals. *J Forensic Sci*. 2020;65(2):465-470. doi:10.1111/1556-4029.14185

72. Boroni M, Zonari A, Reis de Oliveira C, et al. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. *Clin Epigenetics*. 2020;12(1):105. doi:10.1186/s13148-020-00899-1

73. Voisin S, Harvey NR, Haupt LM, et al. An epigenetic clock for human skeletal muscle. *J Cachexia Sarcopenia Muscle*. 2020;11(4):887-898. doi:10.1002/jcsm.12556

74. Belsky DW, Caspi A, Arseneault L, et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. Hagg S, Tyler JK, Hagg S, Justice J, Suderman M, eds. *eLife*. 2020;9:e54870. doi:10.7554/eLife.54870

75. Horvath S, Haghani A, Zoller JA, et al. Pan-primate DNA methylation clocks. Published online November 3, 2021:2020.11.29.402891. doi:10.1101/2020.11.29.402891

76. Jasinska AJ, Haghani A, Zoller JA, et al. Epigenetic clock and methylation studies in vervet monkeys. *GeroScience*. Published online September 30, 2021. doi:10.1007/s11357-021-00466-3

77. Schachtschneider KM, Schook LB, Meudt JJ, et al. Epigenetic clock and DNA methylation analysis of porcine models of aging and obesity. *GeroScience*. 2021;43(5):2467-2483. doi:10.1007/s11357-021-00439-6

78. Prado NA, Brown JL, Zoller JA, et al. Epigenetic clock and methylation studies in elephants. *Aging Cell*. 2021;20(7):e13414. doi:10.1111/acel.13414

79. Raj K, Szladovits B, Haghani A, et al. Epigenetic clock and methylation studies in cats. *GeroScience*. 2021;43(5):2363-2378. doi:10.1007/s11357-021-00445-8

80. Horvath S, Zoller JA, Haghani A, et al. Epigenetic clock and methylation studies in the rhesus macaque. *GeroScience*. 2021;43(5):2441-2453. doi:10.1007/s11357-021-00429-8

81. Horvath S, Zoller JA, Haghani A, et al. DNA methylation age analysis of rapamycin in common marmosets. *GeroScience*. 2021;43(5):2413-2425. doi:10.1007/s11357-021-00438-7

82. Moore GWK, Howell SEL, Brady M, Xu X, McNeil K. Anomalous collapses of Nares Strait ice arches leads to enhanced export of Arctic sea ice. *Nat Commun*. 2021;12:1. doi:10.1038/s41467-020-20314-w

83. Bors EK, Baker CS, Wade PR, et al. An epigenetic clock to estimate the age of living beluga whales. *Evol Appl*. 2021;14(5):1263-1273. doi:10.1111/eva.13195

84. Sugrue VJ, Zoller JA, Narayan P, et al. Castration delays epigenetic aging and feminizes DNA methylation at androgen-regulated loci. *eLife*. 10:e64932. doi:10.7554/eLife.64932

85. Horvath S, Haghani A, Zoller JA, et al. Methylation studies in Peromyscus: aging, altitude adaptation, and monogamy. *GeroScience*. 2021;44(1):447-461. doi:10.1007/s11357-021-00472-5

86. Kordowitzki P, Haghani A, Zoller JA, et al. Epigenetic clock and methylation study of oocytes from a bovine model of reproductive aging. *Aging Cell*. 2021;20(5):e13349. doi:10.1111/acel.13349

87. Steg LC, Shireby GL, Imm J, et al. Novel epigenetic clock for fetal brain development predicts prenatal age for cellular stem cell models and derived neurons. *Mol Brain*. 2021;14:98. doi:10.1186/s13041-021-00810-w

88. Graw S, Camerota M, Carter BS, et al. NEOage clocks - epigenetic clocks to estimate post-menstrual and postnatal age in preterm infants. *Aging*. 2021;13(20):23527-23544. doi:10.18632/aging.203637

89. Galkin F, Mamoshina P, Kochetov K, Sidorenko D, Zhavoronkov A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. *Aging Dis*. 2021;12(5):1252-1262. doi:10.14336/AD.2020.1202

90. Pinho GM, Martin JGA, Farrell C, et al. Hibernation slows epigenetic ageing in yellow-bellied marmots. *Nat Ecol Evol*. 2022;6(4):418-426. doi:10.1038/s41559-022-01679-1

91. Kerepesi C, Meer MV, Ablaeva J, et al. Epigenetic aging of the demographically non-aging naked mole-rat. *Nat Commun*. 2022;13(1):355. doi:10.1038/s41467-022-27959-9

92. Horvath S, Haghani A, Peng S, et al. DNA methylation aging and transcriptomic studies in horses. *Nat Commun*. 2022;13:40. doi:10.1038/s41467-021-27754-y

93. Lemaître JF, Rey B, Gaillard JM, et al. DNA methylation as a tool to explore ageing in wild roe deer populations. *Mol Ecol Resour*. 2022;22(3):1002-1015. doi:10.1111/1755-0998.13533

94. Minteer C, Morselli M, Meer M, et al. Tick tock, tick tock: Mouse culture and tissue aging captured by an epigenetic clock. *Aging Cell*. 2022;21(2):e13553. doi:10.1111/acel.13553

95. Mayne B, Mustin W, Baboolal V, et al. Age prediction of green turtles with an epigenetic clock. *Mol Ecol Resour*. n/a(n/a). doi:10.1111/1755-0998.13621

96. Belsky DW, Caspi A, Corcoran DL, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife*. 2022;11:e73420. doi:10.7554/eLife.73420

97. Liang X, Sinha R, Justice AC, Cohen MH, Aouizerat BE, Xu K. A new monocyte epigenetic clock reveals nonlinear effects of alcohol consumption on biological aging in three independent cohorts (N = 2242). *Alcohol Clin Exp Res*. 2022;46(5):736-748. doi:10.1111/acer.14803

98. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *Npj Aging*. 2022;8(1):1-15. doi:10.1038/s41514-022-00085-y

99. Sugden K, Hannon EJ, Arseneault L, et al. Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. *Patterns*. 2020;1(2):100014. doi:10.1016/j.patter.2020.100014

100. Cao X, Li W, Wang T, et al. Accelerated biological aging in COVID-19 patients. *Nat Commun*. 2022;13(1):2135. doi:10.1038/s41467-022-29801-8

101. Corley MJ, Pang APS, Dody K, et al. Genome-wide DNA methylation profiling of peripheral blood reveals an epigenetic signature associated with severe COVID-19. *J Leukoc Biol*. 2021;110(1):21-26. doi:10.1002/JLB.5HI0720-466R

102. Franzen J, Nüchtern S, Tharmapalan V, et al. Epigenetic Clocks Are Not Accelerated in COVID-19 Patients. *Int J Mol Sci*. 2021;22(17):9306. doi:10.3390/ijms22179306

103. Carroll JE, Ross KM, Horvath S, et al. Postpartum sleep loss and accelerated epigenetic aging. *Sleep Health*. 2021;7(3):362-367. doi:10.1016/j.sleh.2021.02.002

104. Hamlat EJ, Prather AA, Horvath S, Belsky J, Epel ES. Early life adversity, pubertal timing, and epigenetic age acceleration in adulthood. *Dev Psychobiol*. 2021;63(5):890-902. doi:10.1002/dev.22085

105. Nishitani S, Suzuki S, Ochiai K, et al. Altered epigenetic clock in children exposed to maltreatment. *Psychiatry Clin Neurosci*. 2021;75(3):110-112. doi:10.1111/pcn.13183

106. Dammering F, Martins J, Dittrich K, et al. The pediatric buccal epigenetic clock identifies significant ageing acceleration in children with internalizing disorder and maltreatment exposure. *Neurobiol Stress*. 2021;15:100394. doi:10.1016/j.ynstr.2021.100394

107. Roberts JD, Vittinghoff E, Lu AT, et al. Epigenetic Age and the Risk of Incident Atrial Fibrillation. *Circulation*. 2021;144(24):1899-1911. doi:10.1161/CIRCULATIONAHA.121.056456

108. Fitzgerald KN, Hodges R, Hanes D, et al. Potential reversal of epigenetic age using a diet and lifestyle intervention: a pilot randomized clinical trial. *Aging*. 2021;13(7):9419-9432. doi:10.18632/aging.202913

109. Tekola-Ayele F. Invited Commentary: Epigenetic Clocks and Obesity-Towards the Next Frontier Using Integrative Approaches and Early-Life Models. *Am J Epidemiol*. 2021;190(6):994-997. doi:10.1093/aje/kwaa252

110. Fiorito G, Caini S, Palli D, et al. DNA methylation-based biomarkers of aging were slowed down in a two-year diet and physical activity intervention trial: the DAMA study. *Aging Cell*. 2021;20(10):e13439. doi:10.1111/acel.13439

111. Chen Z, Stanbouly S, Nishiyama NC, et al. Spaceflight decelerates the epigenetic clock orchestrated with a global alteration in DNA methylome and transcriptome in the mouse retina. *Precis Clin Med*. 2021;4(2):93-108. doi:10.1093/pcmedi/pbab012

112. Wu X, Ye J, Wang Z, Zhao C. Epigenetic Age Acceleration Was Delayed in Schizophrenia. *Schizophr Bull*. 2021;47(3):803-811. doi:10.1093/schbul/sbaa164

113. Neri de Souza Reis V, Tahira AC, Daguano Gastaldi V, et al. Environmental Influences Measured by Epigenetic Clock and Vulnerability Components at Birth Impact Clinical ASD Heterogeneity. *Genes*. 2021;12(9):1433. doi:10.3390/genes12091433

114. Bertucci EM, Mason MW, Rhodes OE, Parrott BB. Exposure to ionizing radiation disrupts normal epigenetic aging in Japanese medaka. *Aging*. 2021;13(19):22752-22771. doi:10.18632/aging.203624

115. Xiao C, Beitler JJ, Peng G, et al. Epigenetic age acceleration, fatigue, and inflammation in patients undergoing radiation therapy for head and neck cancer: A longitudinal study. *Cancer*. 2021;127(18):3361-3371. doi:10.1002/cncr.33641

116. Fransquet PD, Lacaze P, Saffery R, et al. Accelerated Epigenetic Aging in Peripheral Blood does not Predict Dementia Risk. *Curr Alzheimer Res*. 2021;18(5):443-451. doi:10.2174/1567205018666210823100721

117. Sehl ME, Henry JE, Storniolo AM, Horvath S, Ganz PA. The Effects of Lifetime Estrogen Exposure on Breast Epigenetic Age. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2021;30(6):1241-1249. doi:10.1158/1055-9965.EPI-20-1297

118. Mehta D, Bruenig D, Pierce J, et al. Recalibrating the epigenetic clock after exposure to trauma: The role of risk and protective psychosocial factors. *J Psychiatr Res*. 2022;149:374-381. doi:10.1016/j.jpsychires.2021.11.026

119. Hoare J, Stein DJ, Heany SJ, et al. Accelerated epigenetic aging in adolescents living with HIV is associated with altered development of brain structures. *J Neurovirol*. Published online February 7, 2021. doi:10.1007/s13365-021-00947-3

120. Esteban-Cantos A, Rodríguez-Centeno J, Barruz P, et al. Epigenetic age acceleration changes 2 years after antiretroviral therapy initiation in adults with HIV: a substudy of the NEAT001/ANRS143 randomised trial. *Lancet HIV*. 2021;8(4):e197-e205. doi:10.1016/S2352-3018(21)00006-0

121. Saptarshi N, Green D, Cree A, Lotery A, Paraoan L, Porter LF. Epigenetic Age Acceleration Is Not Associated with Age-Related Macular Degeneration. *Int J Mol Sci*. 2021;22(24):13457. doi:10.3390/ijms222413457

122. Alsaggaf R, Katta S, Wang T, et al. Epigenetic Aging and Hematopoietic Cell Transplantation in Patients With Severe Aplastic Anemia. *Transplant Cell Ther*. 2021;27(4):313.e1-313.e8. doi:10.1016/j.jtct.2021.01.013

123. Kuo PL, Moore AZ, Lin FR, Ferrucci L. Epigenetic Age Acceleration and Hearing: Observations From the Baltimore Longitudinal Study of Aging. *Front Aging Neurosci*. 2021;13:790926. doi:10.3389/fnagi.2021.790926

124. Monasso GS, Jaddoe VWV, Küpers LK, Felix JF. Epigenetic age acceleration and cardiovascular outcomes in school-age children: The Generation R Study. *Clin Epigenetics*. 2021;13(1):205. doi:10.1186/s13148-021-01193-4

125. Lemke E, Vetter VM, Berger N, Banszerus VL, König M, Demuth I. Cardiovascular health is associated with the epigenetic clock in the Berlin Aging Study II (BASE-II). *Mech Ageing Dev*. 2022;201:111616. doi:10.1016/j.mad.2021.111616

126. Martin CL, Ward-Caviness CK, Dhingra R, et al. Neighborhood environment, social cohesion, and epigenetic aging. *Aging*. 2021;13(6):7883-7899. doi:10.18632/aging.202814

127. Qin N, Li Z, Song N, et al. Epigenetic Age Acceleration and Chronic Health Conditions Among Adult Survivors of Childhood Cancer. *J Natl Cancer Inst*. 2021;113(5):597-605. doi:10.1093/jnci/djaa147

128. Schmitz LL, Zhao W, Ratliff SM, et al. The Socioeconomic Gradient in Epigenetic Ageing Clocks: Evidence from the Multi-Ethnic Study of Atherosclerosis and the Health and Retirement Study. *Epigenetics*. Published online July 6, 2021:1-23. doi:10.1080/15592294.2021.1939479

129. Crimmins EM, Thyagarajan B, Levine ME, Weir DR, Faul J. Associations of Age, Sex, Race/Ethnicity, and Education With 13 Epigenetic Clocks in a Nationally Representative U.S. Sample: The Health and Retirement Study. *J Gerontol A Biol Sci Med Sci*. 2021;76(6):1117-1123. doi:10.1093/gerona/glab016

130. Sillanpää E, Heikkinen A, Kankaanpää A, et al. Blood and skeletal muscle ageing determined by epigenetic clocks and their associations with physical activity and functioning. *Clin Epigenetics*. 2021;13(1):110. doi:10.1186/s13148-021-01094-6

131. Murach KA, Dimet-Wiley AL, Wen Y, et al. Late-life exercise mitigates skeletal muscle epigenetic aging. *Aging Cell*. 2022;21(1):e13527. doi:10.1111/acel.13527

132. Protsenko E, Yang R, Nier B, et al. "GrimAge," an epigenetic predictor of mortality, is accelerated in major depressive disorder. *Transl Psychiatry*. 2021;11(1):193. doi:10.1038/s41398-021-01302-0

133. van der Laan L, Cardenas A, Vermeulen R, et al. Epigenetic aging biomarkers and occupational exposure to benzene, trichloroethylene and formaldehyde. *Environ Int*. 2022;158:106871. doi:10.1016/j.envint.2021.106871

134. Kho M, Wang YZ, Chaar D, et al. Accelerated DNA methylation age and medication use among African Americans. *Aging*. 2021;13(11):14604-14629. doi:10.18632/aging.203115

135. Gindin Y, Gaggar A, Lok AS, et al. DNA Methylation and Immune Cell Markers Demonstrate Evidence of Accelerated Aging in Patients with Chronic Hepatitis B Virus or Hepatitis C Virus, with or without Human Immunodeficienct Virus Co-infection. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2021;73(1):e184-e190. doi:10.1093/cid/ciaa1371

136. Matías-García PR, Ward-Caviness CK, Raffield LM, et al. DNAm-based signatures of accelerated aging and mortality in blood are associated with low renal function. *Clin Epigenetics*. 2021;13(1):121. doi:10.1186/s13148-021-01082-w

137. Okazaki S, Kimura R, Otsuka I, et al. Epigenetic aging in Williams syndrome. *J Child Psychol Psychiatry*. Published online April 13, 2022. doi:10.1111/jcpp.13613

138. Cardenas A, Ecker S, Fadadu RP, et al. Epigenome-wide association study and epigenetic age acceleration associated with cigarette smoking among Costa Rican adults. *Sci Rep*. 2022;12:4277. doi:10.1038/s41598-022-08160-w

139. Vetter VM, Sommerer Y, Kalies CH, Spira D, Bertram L, Demuth I. Vitamin D supplementation is associated with slower epigenetic aging. *GeroScience*. Published online May 13, 2022. doi:10.1007/s11357-022-00581-9

140. Dugger DT, Calabrese DR, Gao Y, et al. Lung Allograft Epithelium DNA Methylation Age Is Associated With Graft Chronologic Age and Primary Graft Dysfunction. *Front Immunol*. 2021;12:704172. doi:10.3389/fimmu.2021.704172

141. Cristoferi I, Giacon TA, Boer K, et al. The applications of DNA methylation as a biomarker in kidney transplantation: a systematic review. *Clin Epigenetics*. 2022;14(1):20. doi:10.1186/s13148-022-01241-7

142. Handl L, Jalali A, Scherer M, Eggeling R, Pfeifer N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics*. 2019;35(14):i154-i163. doi:10.1093/bioinformatics/btz338

143. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.

144. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499. doi:10.1214/009053604000000067

145. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Epigenetics*. 2019;11(1):123. doi:10.1186/s13148-019-0730-1

146. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288.

147. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 2000;42(1):80-86. doi:10.1080/00401706.2000.10485983

148. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R (Springer Texts in Statistics)*. 2013th ed. Springer; 2013.

149. Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*. 2012;13(1):59. doi:10.1186/1471-2105-13-59

150. Gupta S, Gupta A. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Comput Sci*. 2019;161:466-474. doi:10.1016/j.procs.2019.11.146

151. The Cancer Genome Atlas Program - NCI. Published June 13, 2018. Accessed June 3, 2022. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

152. Song MA, Brasky TM, Weng DY, et al. Landscape of genome-wide age-related DNA methylation in breast tissue. *Oncotarget*. 2017;8(70):114648-114662. doi:10.18632/oncotarget.22754

153. Charlton J, Williams RD, Sebire NJ, et al. Comparative methylome analysis identifies new tumour subtypes and biomarkers for transformation of nephrogenic rests into Wilms tumour. *Genome Med*. 2015;7(1):11. doi:10.1186/s13073-015-0136-4

154. GDC FAQs | NCI Genomic Data Commons. Accessed December 20, 2021. https://gdc.cancer.gov/about-gdc/gdc-faqs

155. GDC Data Processing | NCI Genomic Data Commons. Accessed June 13, 2022. https://gdc.cancer.gov/about-data/gdc-data-processing

156. GDC Data Transfer Tool | NCI Genomic Data Commons. Accessed June 13, 2022. https://gdc.cancer.gov/access-data/gdc-data-transfer-tool

157. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol Clifton NJ*. 2016;1418:93-110. doi:10.1007/978-1-4939-3578-9_5

158. Cerami1 E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*. 2012;2(5):401-404. doi:10.1158/2159-8290.CD-12-0095

159. Cancer Genomics | cBio@MSKCC. Accessed June 13, 2022. https://cbio.mskcc.org/tools/cancer-genomics/index.html

160. RStudio Team. *RStudio: Integrated Development for R.* RStudio, PBC.; 2022. http://www.rstudio.com/

161. Macintyre G, Naeem H, Wong NC, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics*. 2014;15(1):51. doi:10.1186/1471-2164-15-51

162. Colaprico A, Silva TC, Olsen C, et al. *TCGAbiolinks: TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis with GDC Data*. Bioconductor version: Release (3.14); 2021. doi:10.18129/B9.bioc.TCGAbiolinks

163. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/

164. Ramos M, Geistlinger L, Oh S, et al. Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clin Cancer Inform*. 2020;(4):958-971. doi:10.1200/CCI.19.00119

165. Li A, Kane AE, Mueller A, et al. Novel feature selection methods for construction of accurate epigenetic clocks. Published online February 22, 2022:2022.02.21.481326. doi:10.1101/2022.02.21.481326

166. Charlton J, Williams RD, Weeks M, et al. Methylome analysis identifies a Wilms tumor epigenetic biomarker detectable in blood. *Genome Biol*. 2014;15(8):434. doi:10.1186/s13059-014-0434-y

167. Hofstatter EW, Horvath S, Dalela D, et al. Increased epigenetic age in normal breast tissue from luminal breast cancer patients. *Clin Epigenetics*. 2018;10:112. doi:10.1186/s13148-018-0534-8