PLOS ONE

RESEARCH ARTICLE

RNAAgeCalc: A multi-tissue transcriptional age calculator

Xu Ren, Pei Fen Kuan*

Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, United States of America

* peifen.kuan@stonybrook.edu

Citation: Ren X, Kuan PF (2020) RNAAgeCalc: A multi-tissue transcriptional age calculator. PLoS ONE 15(8): e0237006. https://doi.org/10.1371/journal.pone.0237006

Editor: Yun Li, University of North Carolina at Chapel Hill, UNITED STATES

Received: May 25, 2020

Accepted: July 16, 2020

Published: August 4, 2020

Data Availability Statement: GTEx RNA-Seq data is publicly available at https://www.gtexportal.org/home/datasets under GTEx Analysis V6. This dataset can also be downloaded from https://jhubiostatistics.shinyapps.io/recount/ under the GTEx tab. TCGA clinical, DNA methylation and RNA-Seq data are publicly available at https://gdac.broadinstitute.org/. These datasets can also be downloaded from https://jhubiostatistics.shinyapps.io/recount/ under the TCGA tab. The GTEx chronological age and self-reported race cannot be shared publicly because they are GTEx protected data. These data are available at dbGaP (https://www.ncbi.nlm.nih.gov/gap/) for

Abstract

Biological aging reflects decline in physiological functions and is an effective indicator of morbidity and mortality. Numerous epigenetic age calculators are available, however biological aging calculators based on transcription remain scarce. Here, we introduce RNAAgeCalc, a versatile across-tissue and tissue-specific transcriptional age calculator. By performing a meta-analysis of transcriptional age signature across multi-tissues using the GTEx database, we identify 1,616 common age-related genes, as well as tissue-specific age-related genes. Based on these genes, we develop new across-tissue and tissue-specific age predictors. We show that our transcriptional age calculator outperforms other prior age related gene signatures as indicated by the higher correlation with chronological age as well as lower median and median error. Our results also indicate that both racial and tissue differences are associated with transcriptional age. Furthermore, we demonstrate that the transcriptional age acceleration computed from our within-tissue predictor is significantly correlated with mutation burden, mortality risk and cancer stage in several types of cancer from the TCGA database, and offers complementary information to DNA methylation age. RNAAgeCalc is available at http://www.ams.sunysb.edu/~pfkuan/softwares.html#RNAAgeCalc, both as Bioconductor and Python packages, accompanied by a user-friendly interactive Shiny app.

Introduction

Aging is among the most complex phenotype and is a well-known risk factor for a myriad of diseases including cardiovascular, diabetes, arthritis, neurodegeneration and cancer [1]. Increasing evidence has pointed to the interactions between genetics, epigenetics and environmental factors in the aging process [2]. Over the last decade, there has been a growing body of research in identifying genetic and epigenetic biomarkers of aging to decipher the molecular mechanisms underpinning disease susceptibility. For example, the genome-wide association studies (GWAS) have identified genetic loci associated with longevity and several aging-related diseases [3–6]. As aging is a multifactorial process determined by the dynamic nature of static genetics as well as stochastic epigenetic variation and transcriptomics regulation, both DNA

researchers who meet the criteria for access to confidential data. The authors had no special access privileges and other researchers will be able to access the data in the same manner as the authors. Contact: Data Access Committee for National Human Genome Research Institute (nhgridac@mail.nih.gov).

Funding: PK: CDC/NIOSH award U01 OH011478 https://www.cdc.gov/niosh/awards/default.html The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

methylation and gene expression have emerged as promising hallmark for understanding the aging process and its associated diseases.

Numerous estimators have been developed to predict human aging from DNA methylation data profiled on the Illumina Infinium HumanMethylation450K BeadChip. Among the most widely used epigenetic age estimator are the DNA methylation age calculator of Horvath [7] and Hannum et al [8], by regressing chronological age on DNA methylation. Specifically, Horvath [7] derived a multi-tissue and cell types DNA methylation age (DNAm age) estimator across the entire lifespan of human. The predictor was constructed using > 8,000 non-cancer samples from 82 DNA methylation datasets, including 51 healthy tissues and cell types based on elastic net [9] model, a penalized regression statistical framework which retained 353 CpGs in the final model. On the other hand, Hannum et al [8] developed a 71-CpGs DNA methylation age calculator based on the whole blood of 656 human samples aged 19 to 101. While the first generation DNA methylation age estimators including Horvath's clock and Hannum's clock were developed based on chronological age, the second generation DNA methylation age estimators were obtained by optimizing the prediction error on phenotypic age derived from clinical attributes associated with mortality and morbidity. This includes PhenoAge [10] and GrimAge [11] which aim to improve prediction of aging related outcomes (e.g., time-to-death, time-to-disease for cancer, Alzheimer's disease and cardiovascular).

In addition to DNA methylation, changes in gene expression have been shown to be associated with aging and aging-related outcomes [12–19]. Specifically, de Magalhaes et al. [12] identified 56 and 17 genes consistently over- and under-expressed with chronological age, respectively by performing a meta-analysis on 27 microarray datasets from mice, rats and human subjects. Their age specific signature was obtained by first regressing logarithm transformed gene expression on chronological age for each individual microarray dataset. The significance of differential expression in each dataset was determined via a two-tailed F-test, followed by binomial tests to identify genes that were consistently over- or under-expressed across datasets. This study was based on microarray data and 23 out of 27 datasets were from mice or rats subjects, potentially limiting the transferability of the derived age related gene signature to human datasets. As shown in the Results section below, the gene signature of de Magalhaes's resulted in biased prediction of RNA age based on human RNA-Seq datasets. A closely related work was the development of the GenAge (version 19) database of aging-related genes, including 307 genes potentially related to human aging [13]. Unlike majority of the gene signatures which were typically derived from statistical models (e.g., study specific association analysis or meta-analysis), the genes in GenAge were manually curated by summarizing the biological properties of genes from > 2,000 references on human and animal aging studies across different tissues.

Besides the de Magalhaes et al. [12] signature and GenAge which included across-tissue age-related genes, there were several gene signatures derived for individual tissues. For example, Welle et al. [14] identified 718 probe sites that were related to aging in human muscle using the microarray data from 8 healthy young men and 8 healthy old men. Rodwell et al. [15] identified 985 genes related to aging in human kidney by analyzing the microarray data from 74 healthy kidney with age ranging from 27 to 92 years old, whereas Lu et al. [16] identified 463 aging-related genes in human brain by analyzing the microarray data from 30 samples with age ranging from 26 to 106 years old. Glass et al. [17] identified 1,672 and 188 genes associated with age in skin and adipose tissue, respectively from 856 female twin samples profiled on Illumina Human HT-12 V3 Bead chip.

To the best of our knowledge, the largest meta-analysis to identify age related genes was conducted by Peters et al. [18] from whole blood gene expression of 14,983 human subjects of European ancestry, profiled using microarray platform Illumina Human HT-12 V3 and HT-

12 V4 BeadChip. 7,074 samples were used to construct the gene expression signature whereas the remaining samples were used to test the signature. A total of 1,497 genes was significant in both the training and test dataset. The authors further showed that the differences between the predicted transcriptional age and chronological age were associated with biological features related to aging, including blood pressure, cholesterol levels, fasting glucose and body mass index. On the other hand, the most recent transcriptional age predictor was developed by Fleischer et al. [19] based on a novel ensemble linear discriminant analysis (LDA) method using human dermal fibroblast data. Their dataset consisted of gene expression profiled using RNA-Seq from 133 healthy samples with age ranging from 1 to 94 years. The authors further showed that ensemble LDA outperformed other prediction algorithms including elastic net, support vector regression, and linear regression in terms of mean/median absolute difference between predicted age and chronological age. By using leave-one-out cross validation, the authors were able to obtain 4 years median absolute error and 7.8 years mean absolute error in their dermal fibroblast dataset.

Unlike DNA methylation in which several user-friendly software and computer programs are available for predicting epigenetic age across different tissues on the most widely utilized Illumina methylation platform, there were limited transcriptional age predictors and the existing predictors have several pitfalls. First, most of the human transcriptional age predictors were developed based on microarray data and/or limited to only a few tissues. Second, the only predictor constructed using RNA-Seq data was the ensemble LDA predictor [19]. However, this predictor was derived based only on fibroblast data. To date, transcriptional studies on aging using RNA-Seq data across different human tissues was limited. Recognizing the gap in existing research of transcriptional aging based on RNA-Seq data, the aim of this study was twofold, first to identify common age-related genes across tissues; second to construct tissue-specific transcriptional age calculators for understanding how gene expression changed with age in different human tissues. To this end, we utilized a large publicly available RNA-Seq datasets as described in the following sections and developed a transcriptional age predictor for RNA-Seq data. Our transcriptional age predictor is available both as Bioconductor and Python packages RNAAgeCalc, accompanied by a user-friendly interactive Shiny app.

Results

An overview of the transcriptional age analysis pipeline and comparisons conducted in this study was given in Fig 1.

Meta-analysis of age signature across multi-tissues public RNA-Seq dataset

We utilized the RNA-Seq data from the Genotype-Tissue Expression (GTEx) Program [20], a publicly available database to identify across-tissue genes and construct our tissue-specific transcriptional age calculator. We used GTEx V6 release which contained gene expression data at gene, exon, and transcript level of 9,662 samples across 30 different tissues. Since tumor showed notably different gene expression patterns compared to non-tumor [21], the 102 tumor samples from GTEx V6 release were omitted. The remaining samples with complete gender and race information were used in the subsequent analysis. A list of GTEx tissues and summary of non-tumor sample size by gender and race was given in S1 Table. The data preprocessing steps were described in the Methods section.

One pitfall of individual gene expression studies for identifying age-related signatures was the low overlap between the gene lists across different tissues [17, 22]. Meta-analysis of gene expression for aggregating information from different datasets is a useful approach to identify weak genetic signals and has been shown to be powerful in cancer studies [23]. In this paper,

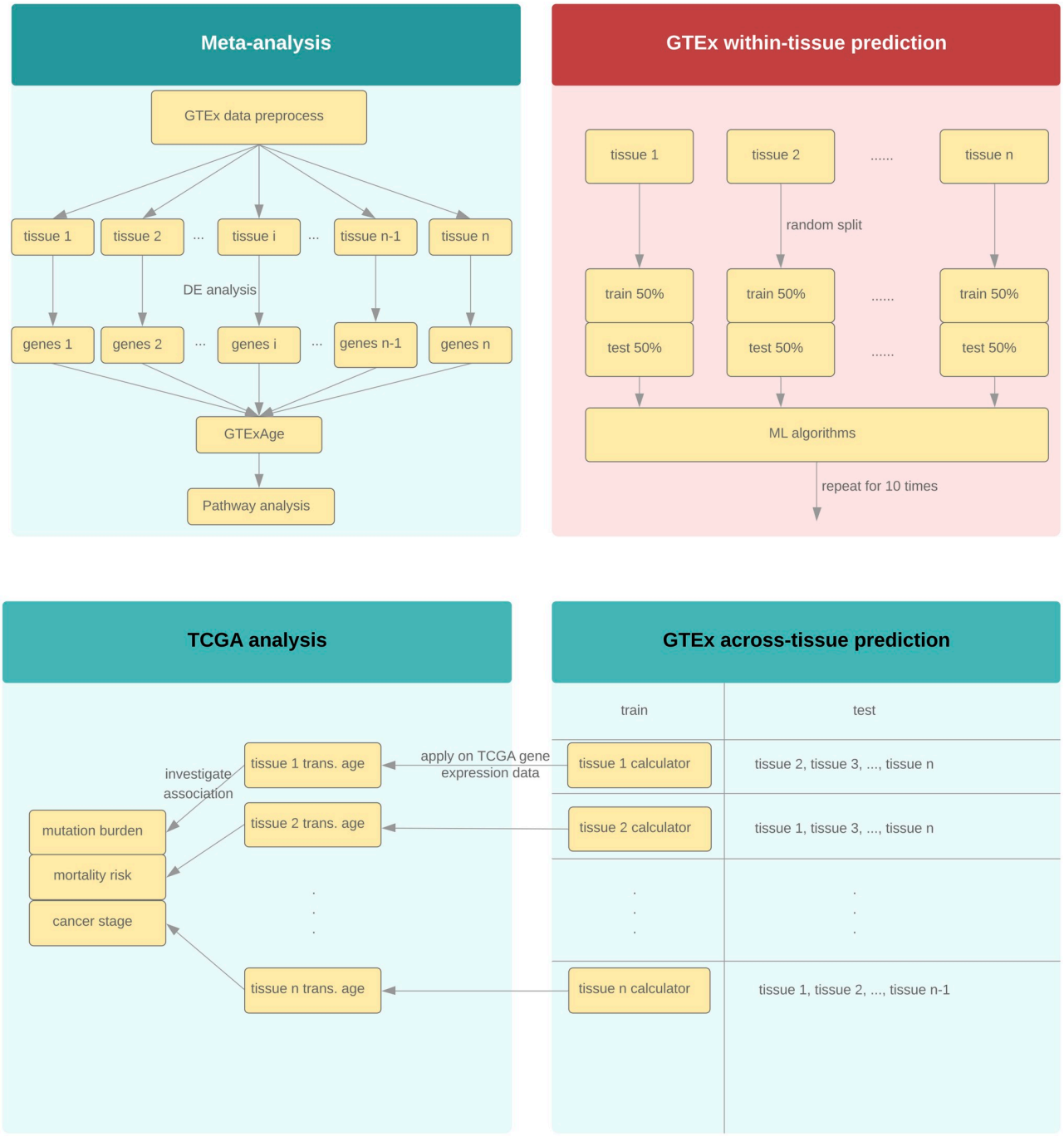

Fig 1. Overview of the transcriptional age analysis pipeline.

https://doi.org/10.1371/journal.pone.0237006.g001

we performed a meta-analysis on the different tissues from the GTEx datasets, aiming at identifying common age-related gene expression signatures across the different tissues. After preprocessing, a total of 26 tissues across 9,448 samples were included in our analysis. A summary of number of genes, number of significant genes associated with age under different FDR cutoffs was provided in S2 Table. Certain tissues (e.g. colon, brain) showed strong signal with a high proportion of differentially expressed genes whereas other tissues including small

intestine, pancreas, pituitary exhibited relatively weak signal, with only a small proportion of differentially expressed genes. At FDR < 0.05, we obtained 0.02%—33.08% genes with positive association and 0%—36.09% genes with negative association across tissues. On average, 13.91% and 13.85% genes were positively and negatively associated with age, respectively. The differentially expressed genes had little overlap across tissues. Among the 26 tissues analyzed, no gene was common across all tissues. Only one gene (EDA2R) was differentially expressed in at least 20 tissues, supporting that age-related signatures were tissue-specific.

To overcome the low overlap across tissues and to identify common age-related genes across tissues, we adapted the binomial test of de Magalhaes et al. [12]. A total of 1,616 common age-related gene across tissues (gender and race adjusted) were identified at FDR < 0.05, as listed in S3 Table. These 1,616 genes are referred to as GTExAge thereafter. The details of our approach was described in the Methods section.

Gene sets associated with common age-related genes

The list of genes which exhibited consistent positive association with age were enriched in GO terms related to plasma-membrane adhesion molecules, response to interferon-gamma, GTPase activity, and type I interferon. The enrichment analysis of genes negatively associated with age identified KEGG terms including proteasome, ribosome biogenesis, RNA transport in eukaryotes, citrate cycle, carbon metabolism, pyruvate metabolism, aminoacyl-tRNA biosynthesis as well as GO terms related to mitochondrial function, metabolic process, RNA processing, ribosome biogenesis, and purine metabolic process. Our results were consistent with the findings of Peters et al. [18] which showed that genes involved in RNA metabolism, ribosome biogenesis, purine metabolism, mitochondrial and metabolic pathways were negatively correlated with age. In addition, genes involved in metabolism and mitochondrial protein synthesis were shown to be down regulated in muscle [14]. The study of aging in human brain [16] also indicated that age-related genes in brain played central roles in mitochondrial function. A full list of significant KEGG and GO terms was provided in S3 Table.

Within-tissue age prediction and tissue-specific age-related genes

Previous studies showed that age-related signatures were tissue-specific [22]. In this paper, we obtained similar conclusion for the GTEx datasets (see meta-analysis section). To assess whether RNA-Seq data was able to predict chronological age accurately, we first evaluated the age prediction within tissue. We considered a rich class of machine learning prediction models, including elastic net [9], generalized boosted regression models (GBM) [24], random forest [25], support vector regression (SVR) with radial kernel [26], and ensemble LDA [19], as well as numerous candidate feature sets for each algorithm as described in the Methods section. A summary of each candidate feature set and number of genes in each set was provided in S4 Table. Our objective here was twofold, first to evaluate which machine learning algorithm performed the best in terms of age prediction, and second to evaluate whether the within-tissue candidate feature sets had better performance compared to across-tissue candidate feature sets.

S5 Table summarized the average Pearson correlation, Spearman correlation, median error, mean error comparing the predicted age to chronological age for each prediction algorithm, tissue and candidate feature set, across 10 repetitions. Elastic net outperformed other algorithms as illustrated by the highest correlation and lowest mean and median error for almost all tissues and all candidate feature sets. Although ensemble LDA [19] was developed using dermal fibroblast samples, the prediction accuracy on the GTEx fibroblast data was lower than the other algorithms. The prediction accuracy of this model on other tissues was also lower

than the other algorithms. These observations indicated that the ensemble LDA [19] predictor may not be generalizable to predict transcriptional age in other tissues or datasets. As elastic net outperformed the other algorithms, we focused on the age prediction using elastic net in the subsequent analysis. Further comparisons of elastic net and ensemble LDA were provided in S1 Appendix and S6 Table.

In the elastic net model, the tissue-specific candidate feature sets based on DESeq2 and all genes outperformed the across-tissue candidate feature sets in most tissues. Another tissue-specific candidate feature set, namely Peters et al. [18] signature developed using microarray blood samples, performed well on the blood samples in GTEx data. On the other hand, our across-tissue signature GTExAge performed best in adrenal gland, breast, lung, small intestine, and stomach but lower performance compared to tissue-specific candidate feature sets in other tissues. Another across-tissue candidate feature set, namely de Magalhaes et al. [12] signature had low prediction accuracy compared to other signatures, which was partially attributed to the fact that it was developed using a large proportion of non-human samples. Taken together, these results suggested that tissue-specific candidate feature sets performed better than across-tissue candidate feature sets in terms of within-tissue age prediction.

Across-tissue age prediction

To better understand the generalizability of individual tissue-specific transcriptional age predictors to other tissues, we investigated the performance of across-tissue age prediction. Elastic net model was trained on samples from one tissue, and tested on the other tissues. Each gene signature discussed in the Methods section was applied and the Pearson correlation between predicted age and chronological age in test tissue was calculated. The correlation matrix was provided in S7 Table, where the row represented the training tissue whereas the column represented the test tissue. For each predictor, we calculated the average Pearson correlation across the tissues tested. We then averaged the correlation across all the predictors to evaluate the performance of each candidate feature set. Table 1 showed the average correlation and weighted average correlation by sample size. Our across-tissue feature set GTExAge had the highest correlation, followed by the tissue-specific feature set based on all genes. These results suggested that our across-tissue feature set GTExAge was better than tissue-specific feature sets in across-tissue age prediction.

The heatmap of Pearson correlation between predicted age and chronological age in each tissue based on GTExAge candidate feature was given in Fig 2. The heatmaps of all genes and DESeq2 candidate features were given in S1 Fig. Pairs of tissues showing higher correlation were partially attributable to the tissue lineages and functional similarities [27]. For example, transcriptional age predictor trained on adipose tissue predicted transcriptional age in blood

Table 1. Performance evaluation based on average Pearson correlations across tissues for each candidate feature set.

signature	correlation	signature	correlation (sample size adjusted)
GTExAge	0.3266	GTExAge	0.3556
all	0.2910	all	0.3389
DESeq2	0.2524	Pearson	0.2883
Pearson	0.2507	DESeq2	0.2858
Dev	0.2257	Dev	0.2657
Peters [18]	0.1547	Peters [18]	0.1834
GenAge [13]	0.1247	GenAge [13]	0.1272
deMagalhaes [12]	0.0841	deMagalhaes [12]	0.0962

https://doi.org/10.1371/journal.pone.0237006.t001

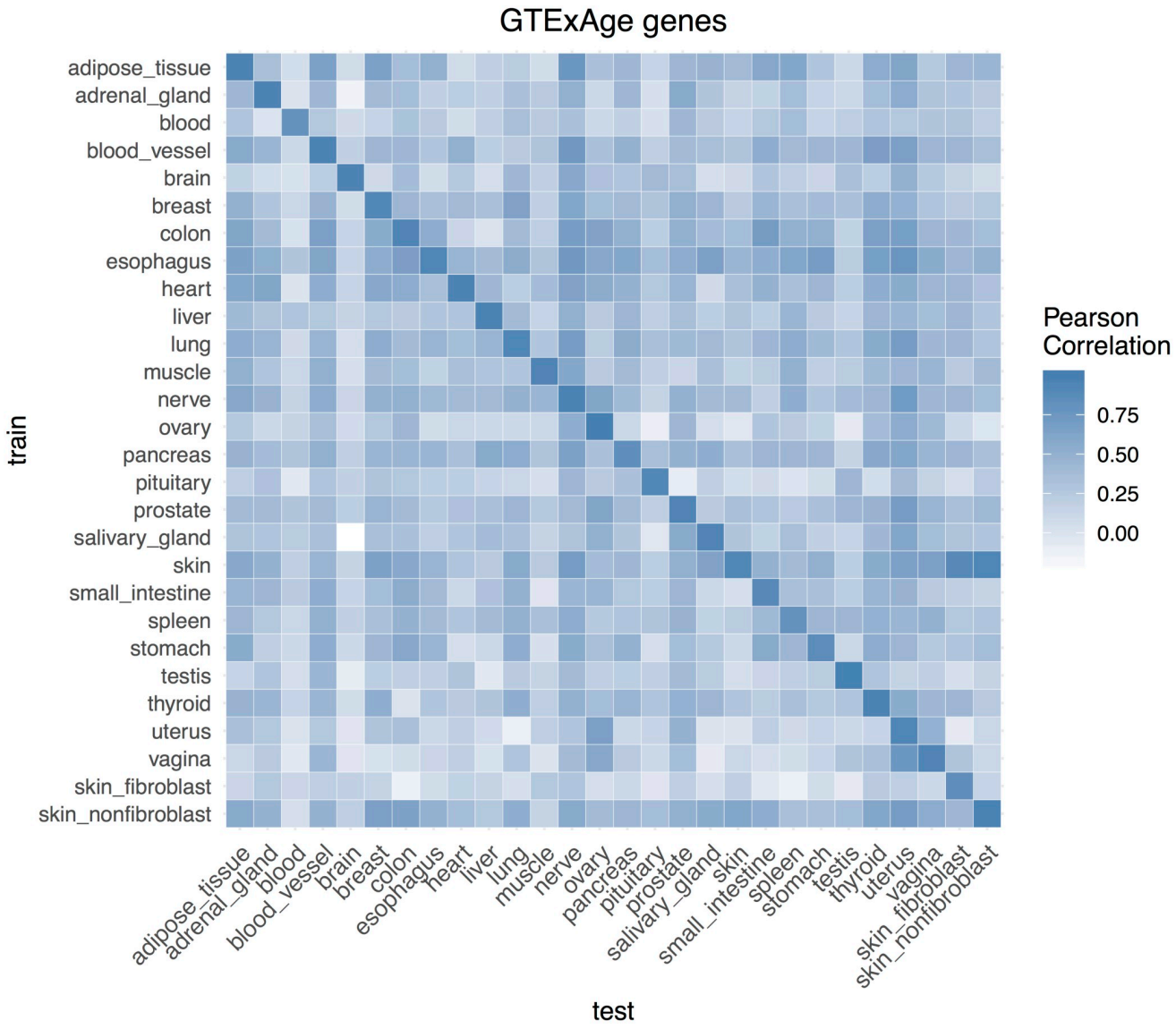

Fig 2. Heat-map of Pearson correlation matrix between predicted age and chronological age (based on GTExAge genes).

https://doi.org/10.1371/journal.pone.0237006.g002

vessel with high correlation ($r = 0.65$ in GTExAge signature) due to the similarity in anatomic and function of these two tissues [28].

Racial effect on transcriptional age predictor

To evaluate the effect of race on age prediction, we repeated the analysis of the meta-analysis, within-tissue prediction, and across-tissue prediction using only subsets of White samples which made up the racial majority in GTEx database. The predictors were trained on the White samples only and tested on White and non-White samples respectively.

For the within-tissue prediction, the White samples were also divided into 50%-50% training-test set. The predictors were built on the training set and evaluated on the test set as well as on the non-White samples. S8 Table summarized average Pearson correlation, Spearman correlation, median error and mean error in each tissue. The predictors tested on White samples

had higher correlation and lower error compared to non-White samples in almost all tissues, which indicated the age-related genes were associated with racial differences.

For the across-tissue prediction, elastic net model was constructed using White samples in each tissue. The model was then tested on the White and non-White samples of all the other tissues respectively. The correlation matrix on the White and non-White samples (based on GTExAge candidate feature) was provided in S9 Table. S2 Fig compared the average correlation on the test sets across all the predictors. The Wilcoxon rank sum test indicated that the difference in the average correlation between the two groups was significant ($p < 0.05$).

Both the within-tissue and across-tissue prediction suggested that transcriptional age predictor was racial-dependent. Thus, in our transcriptional age calculator, we provided the option for computing transcriptional age based on models trained on GTEx White samples only.

Associations with prior aging candidate genes

Here we investigated the intersection between our transcriptional age signature to candidate genes identified from previous studies. First, we compared our tissue-specific signature to prior tissue-specific signatures as summarized in S3 Fig, S10 and S11 Tables.

In general, our tissue specific signatures obtained from DESeq2 analyses were consistent with previous tissue-specific candidate age signatures, except for the comparison with Glass et al. [17] skin signature, which could be attributed to the fact that the skin samples were taken from different anatomic regions. Specifically, the skin samples of Glass et al. [17] were from infra-umbilical whereas the skin samples in GTEx were from suprapubic skin, leg and fibroblast. To investigate the performance of transcriptional age prediction based on these prior aging candidate genes, we performed within-tissue transcriptional age prediction using each of these signatures. The procedure was similar to the within-tissue prediction section except that the prior aging candidate genes were used to train the elastic net model. To evaluate the prediction accuracy, we also reported the root mean squared error (RMSE) of the predicted age. S12 Table summarized the comparison of prediction accuracy between DESeq2/Pearson gene set and the prior candidate genes. For most tissues, the mean/median error and RMSE of our tissue-specific genes selected by DESeq2 or Pearson correlation were lower compared to the prior candidate genes.

Next, we compared our across-tissue signature (GTExAge) to the prior across-tissue signatures, namely de Magalhaes [12], GenAge [13] and Horvath [7] signatures. Horvath [7] signature was based on 353 CpGs from DNA methylation, thus we considered the genes mapping to these 353 CpGs in our comparison. We compared each of these prior across-tissue signature to GTExAge signature by investigating their p-values in the binomial test for identifying common age-related genes across tissues (see Methods section for details). For each gene signature, only a small proportion of genes were significant at $p < 0.05$ (S13 Table), indicating that our signature GTExAge provided additional insights into age-related genes.

Association between transcriptional age and mutation burden in cancer

Since both methylome and gene expression play important roles in aging, to assess whether they complement each other, we compared age-associated methylation to age-associated gene expression in different tissues. We utilized the DNA methylation data and gene expression data of The Cancer Genome Atlas (TCGA) consortium, a rich repository consisting of omics data for multiple types of cancers. The samples with matched DNA methylation data and RNASeq data were analyzed (see Methods section).

The number of mutations per cancer sample (mutation burden) was previously shown to be negatively correlated with DNAm age acceleration [7]. Here, we aimed to determine whether

Table 2. Correlation between age acceleration residual and mutation burden.

	transcriptional age acceleration (based on all genes)					
	Pearson_r	Pearson_pv	Pearson_pvadj	Spearman_r	Spearman_pv	Spearman_pvadj
ACC	-2.63E-01	1.91E-02	7.15E-02	-2.68E-01	1.71E-02	8.56E-02
BRCA	1.35E-02	7.27E-01	8.51E-01	2.48E-02	5.19E-01	7.08E-01
GBMLGG	-2.53E-01	1.93E-09	1.45E-08	-1.90E-01	7.34E-06	5.51E-05
COADREAD	-1.01E-02	8.43E-01	8.51E-01	-3.34E-03	9.48E-01	9.98E-01
ESCA	1.89E-01	1.06E-02	5.28E-02	9.60E-02	1.97E-01	3.70E-01
LIHC	-5.62E-02	2.96E-01	4.44E-01	-4.80E-02	3.72E-01	5.58E-01
LUAD	-2.41E-01	1.00E-11	1.51E-10	-2.43E-01	6.11E-12	9.16E-11
OV	-6.22E-01	7.38E-02	2.21E-01	-5.83E-01	1.08E-01	3.24E-01
PAAD	2.40E-02	7.62E-01	8.51E-01	1.86E-04	9.98E-01	9.98E-01
PRAD	7.15E-02	1.18E-01	2.95E-01	7.51E-02	1.00E-01	3.24E-01
SKCM (tumor)	1.43E-01	1.49E-01	3.20E-01	1.38E-01	1.64E-01	3.51E-01
STAD	4.60E-02	3.84E-01	5.23E-01	5.88E-02	2.66E-01	4.43E-01
TGCT	1.67E-02	8.51E-01	8.51E-01	-4.39E-02	6.21E-01	7.17E-01
THCA	-5.19E-02	2.56E-01	4.26E-01	-2.39E-02	6.00E-01	7.17E-01
SKCM (metastatic)	6.31E-02	2.32E-01	4.26E-01	7.63E-02	1.49E-01	3.51E-01
all tissues	-1.24E-02	3.76E-01		-5.64E-03	6.88E-01	
	DNAm age acceleration					
	Pearson_r	Pearson_pv	Pearson_pvadj	Spearman_r	Spearman_pv	Spearman_pvadj
ACC	-1.19E-01	2.95E-01	3.65E-01	-9.89E-02	3.86E-01	4.95E-01
BRCA	-1.50E-01	1.00E-04	7.52E-04	-2.32E-01	1.29E-09	1.94E-08
GBMLGG	8.42E-02	4.88E-02	8.13E-02	5.09E-02	2.34E-01	3.51E-01
COADREAD	2.35E-01	3.47E-06	5.21E-05	8.57E-02	9.47E-02	2.03E-01
ESCA	-4.57E-03	9.51E-01	9.51E-01	7.75E-04	9.92E-01	9.92E-01
LIHC	-1.20E-01	2.42E-02	6.06E-02	-1.41E-01	8.12E-03	3.81E-02
LUAD	-1.04E-01	3.82E-03	1.91E-02	-9.24E-02	1.02E-02	3.81E-02
OV	7.85E-02	8.41E-01	9.01E-01	-3.33E-02	9.48E-01	9.92E-01
PAAD	1.41E-01	7.24E-02	1.09E-01	1.19E-01	1.29E-01	2.42E-01
PRAD	-1.05E-01	2.26E-02	6.06E-02	-1.85E-01	5.60E-05	4.20E-04
SKCM (tumor)	1.64E-01	9.69E-02	1.32E-01	1.29E-01	1.93E-01	3.22E-01
STAD	1.27E-01	1.67E-02	6.06E-02	4.51E-02	3.96E-01	4.95E-01
TGCT	-1.88E-01	3.26E-02	6.11E-02	-1.73E-01	5.02E-02	1.51E-01
THCA	4.58E-02	3.16E-01	3.65E-01	1.67E-02	7.15E-01	8.25E-01
SKCM (metastatic)	1.17E-01	2.89E-02	6.11E-02	9.53E-02	7.42E-02	1.85E-01
all tissues	-1.64E-02	2.44E-01		-6.30E-02	7.39E-06	

https://doi.org/10.1371/journal.pone.0237006.t002

transcriptional age acceleration had significant association with mutation burden. To this end, we calculated the number of somatic mutations for each cancer sample in TCGA dataset. For each cancer type, the Pearson and Spearman correlation between age acceleration residual and logarithmic number of mutations were calculated. Correlation tests were performed and the p-values were adjusted via the Benjamini & Hochberg false discovery rate (FDR) [29] procedure (denoted padj). Table 2 compared DNAm age acceleration to transcriptional age acceleration, where the transcriptional age was computed using all genes candidate feature set. The results based on DESeq2 and GTExAge signature as candidate feature set were given in S14 Table. For the transcriptional age acceleration, significant negative associations with mutation burden were observed in brain (GBMLGG) and lung (LUAD) cancer (padj <0.05). Marginal negative

association (padj <0.1) was observed in adrenal gland (ACC). For DNAm age acceleration, negative associations were significant in breast (BRCA), liver (LIHC), lung (LUAD) and prostate (PRAD) cancer. The results indicated that the transcriptional age acceleration and DNAm age acceleration provided complementary information in mutation burden analysis.

Association between transcriptional age and mortality in cancer

We evaluated whether transcriptional age was significantly associated with mortality risk in TCGA datasets. For each cancer type, two Cox proportional hazards models were fitted, namely the Cox regression on age acceleration adjusting for chronological age (Mod0a) and Cox regression on age acceleration adjusting for chronological age, stage, gender and race (Mod1a). In Mod1a, gender was not adjusted for breast (BRCA), ovary (OV), prostate (PRAD) and testis (TGCT) cancer. Table 3 showed that transcriptional age was significantly associated with mortality (padj <0.05 in Mod1a) in brain (GBMLGG) cancer whereas DNAm age was significantly associated with mortality (padj <0.05 in Mod1a) in brain (GBMLGG) and skin (SKCM) metastatic cancer. The association between mortality and transcriptional age acceleration showed consistent effect size direction between Mod0a and Mod1a, and vice versa for the association between mortality and DNAm age acceleration. S15 Table provided the results of transcriptional age constructed using other candidate feature sets.

We further evaluated the association between age acceleration and cancer stage. Two linear regression models were fitted for each cancer type, namely regressing age acceleration on stage adjusting for chronological age (Mod0b) and regressing age acceleration on stage adjusting for chronological age, gender and race (Mod1b). Transcriptional age acceleration was marginally associated with stage (padj <0.1 in Mod1b, Table 4 and S16 Table) in adrenal gland (ACC) and liver (LIHC) cancer, whereas DNAm age acceleration was significantly associated with stage (padj <0.05 in Mod1b) in pancreatic (PAAD) and testicular germ cell (TGCT) cancer.

TCGA matched tumor and normal samples

We applied our tissue-specific predictors on the matched tumor and normal samples from TCGA. Paired t-test and Wilcox test were performed to compare the transcriptional age acceleration residual between tumor and matched normal samples (Fig 3). The tumor samples from breast (BRCA), colon (COADREAD), esophagus (ESCA), prostate (PRAD) and stomach (STAD) cancer showed significant age acceleration (padj <0.05) compared to their matched normal samples. We then investigated the aging rate in these paired samples, which was defined as the ratio of transcriptional age to chronological age. As shown in S4 Fig, the aging rate was significantly higher in tumor samples compared to matched normal samples (padj <0.05) in breast (BRCA), colon (COADREAD), esophagus (ESCA), prostate (PRAD) and stomach (STAD) cancer. On the other hand, the aging rate was lower in liver (LIHC) and thyroid (THCA). Since the second generation DNAm age calculator (PhenoAge [10] and GrimAge [11]) were developed to improve prediction of aging related outcomes, we also recomputed the transcriptional age using the genes corresponding to the CpGs of these calculators. The results were provided in S2 Appendix. Overall, DNAm age acceleration based on PhenoAge showed age acceleration in tumor across all cancer types, whereas the transcriptional age computed based on the genes corresponding to these CpGs showed weaker acceleration pattern. Accelerated aging could signify aberrant chromatin conformation and instability, and represents an early event of malignant transformation of cells [30, 31]. On the other hand, based on our earlier results that some cancer showed negative association between mutation burden, mortality with transcriptional age acceleration, we hypothesized that age associated changes in transcriptome could prevent tumor formation by creating an antiproliferative

Table 3. Coefficient and p-value of age acceleration residual from Cox regression.

	transcriptional age acceleration (based on all genes)					
	Coef_Mod0a	PV_Mod0a	PVadj_Mod0a	Coef_Mod1a	PV_Mod1a	PVadj_Mod1a
ACC	-4.08E-02	1.91E-02	1.02E-01	-1.81E-02	3.34E-01	5.91E-01
BRCA	4.38E-03	5.27E-01	6.20E-01	1.96E-03	7.83E-01	9.03E-01
GBMLGG	-4.18E-02	3.23E-08	4.84E-07	-4.18E-02	3.23E-08	4.84E-07
COADREAD	1.36E-02	7.69E-02	2.00E-01	2.62E-03	7.68E-01	9.03E-01
ESCA	4.95E-03	5.37E-01	6.20E-01	4.89E-03	5.70E-01	7.77E-01
LIHC	-9.36E-03	9.33E-02	2.00E-01	-4.90E-03	3.94E-01	5.91E-01
LUAD	-5.26E-03	9.34E-02	2.00E-01	-3.21E-03	3.14E-01	5.91E-01
OV	-3.76E-02	3.77E-01	5.14E-01	-3.76E-02	3.77E-01	5.91E-01
PAAD	-2.28E-02	2.03E-02	1.02E-01	-2.41E-02	1.96E-02	1.47E-01
PRAD	6.61E-03	9.50E-01	9.50E-01	6.61E-03	9.50E-01	9.50E-01
SKCM (tumor)	-1.12E-02	1.80E-01	3.38E-01	-1.15E-02	2.21E-01	5.91E-01
STAD	-5.57E-04	8.18E-01	8.76E-01	3.46E-04	8.89E-01	9.50E-01
TGCT	4.68E-02	3.76E-01	5.14E-01	5.13E-02	3.77E-01	5.91E-01
THCA	-3.33E-02	2.76E-01	4.60E-01	-3.38E-02	3.13E-01	5.91E-01
SKCM (metastatic)	-6.33E-03	5.23E-02	1.96E-01	-5.67E-03	8.18E-02	4.09E-01
	DNAm age acceleration					
	Coef_Mod0a	PV_Mod0a	PVadj_Mod0a	Coef_Mod1a	PV_Mod1a	PVadj_Mod1a
ACC	-2.66E-02	5.61E-02	1.20E-01	-2.14E-02	1.03E-01	2.57E-01
BRCA	-1.12E-02	4.95E-02	1.20E-01	-7.01E-03	2.22E-01	4.16E-01
GBMLGG	-1.62E-02	1.33E-06	2.00E-05	-1.62E-02	1.33E-06	2.00E-05
COADREAD	-2.60E-04	9.72E-01	9.72E-01	3.07E-03	6.52E-01	8.89E-01
ESCA	-2.23E-02	4.14E-02	1.20E-01	-2.14E-02	6.91E-02	2.50E-01
LIHC	-8.01E-04	8.84E-01	9.47E-01	-1.62E-03	7.77E-01	8.89E-01
LUAD	-8.04E-03	3.26E-02	1.20E-01	-6.51E-03	8.34E-02	2.50E-01
OV	2.18E-02	7.69E-01	9.00E-01	2.18E-02	7.69E-01	8.89E-01
PAAD	-3.82E-03	5.16E-01	8.61E-01	-1.21E-03	8.47E-01	8.89E-01
PRAD	-1.94E-01	1.24E-01	2.33E-01	-1.94E-01	1.24E-01	2.66E-01
SKCM (tumor)	6.48E-03	6.47E-01	9.00E-01	4.62E-03	7.38E-01	8.89E-01
STAD	-1.16E-02	2.42E-02	1.20E-01	-1.20E-02	1.93E-02	9.66E-02
TGCT	1.95E-02	7.50E-01	9.00E-01	7.08E-02	4.10E-01	6.83E-01
THCA	5.01E-03	7.80E-01	9.00E-01	2.53E-03	8.89E-01	8.89E-01
SKCM (metastatic)	-1.24E-02	2.25E-03	1.69E-02	-1.27E-02	1.94E-03	1.46E-02

https://doi.org/10.1371/journal.pone.0237006.t003

barrier for aging cells, akin to a double-edged sword [30]. Although inverse correlation between promoter methylation and gene expression has been observed in different tissues [32], the exact regulatory role of methylation on transcriptome has yet to be uncovered [33]. In particular, transcriptional regulation by methylation in cancer has been shown to be a complex molecular mechanism, characterized by the intricate interplay between SNPs, transcription factors and DNA methylation in regulating gene expression [34–36].

Discussion

DNA methylation and gene expression were associated with aging and aging-related diseases. A number of calculators to predict DNAm age from human DNAm data profiled on the Illumina Infinium HumanMethylation450K BeadChip have been developed. For gene expression data, although several common age-related genes across tissues as well as tissue-specific signatures

Table 4. Coefficient and p-value of age acceleration residual versus stage from linear model.

	transcriptional age acceleration (based on all genes)					
	Coef_Mod0b	PV_Mod0b	PVadj_Mod0b	Coef_Mod1b	PV_Mod1b	PVadj_Mod1b
ACC	-4.0291	0.0123	0.0740	-4.1381	0.0111	0.0668
BRCA	0.3976	0.0678	0.1627	0.3967	0.0673	0.2018
COADREAD	0.4338	0.0320	0.1279	0.4395	0.0290	0.1158
ESCA	0.5447	0.4358	0.5404	0.9255	0.1817	0.3634
LIHC	-1.6087	0.0040	0.0485	-1.5365	0.0057	0.0668
LUAD	-0.5438	0.0638	0.1627	-0.4834	0.0998	0.2396
PAAD	-0.9752	0.2590	0.4440	-0.9918	0.2507	0.3760
SKCM (tumor)	-0.6880	0.4953	0.5404	-0.3771	0.7085	0.7729
STAD	-0.7152	0.4113	0.5404	-0.7701	0.3793	0.5057
TGCT	0.2285	0.2410	0.4440	0.2267	0.2482	0.3760
THCA	-0.1152	0.7951	0.7951	-0.0855	0.8472	0.8472
SKCM (metastatic)	-0.2637	0.4758	0.5404	-0.2707	0.4679	0.5615
	DNAm age acceleration					
	Coef_Mod0b	PV_Mod0b	PVadj_Mod0b	Coef_Mod1b	PV_Mod1b	PVadj_Mod1b
ACC	-0.7054	0.7463	0.7463	-0.5800	0.7924	0.7924
BRCA	-0.2340	0.4224	0.7070	-0.2327	0.4227	0.6853
COADREAD	-0.3468	0.1271	0.4505	-0.3478	0.1271	0.5081
ESCA	-0.3463	0.5302	0.7070	-0.3567	0.5108	0.6853
LIHC	0.4889	0.4480	0.7070	0.4931	0.4428	0.6853
LUAD	-0.3406	0.2192	0.5261	-0.2921	0.2931	0.6853
PAAD	-3.4878	0.0077	0.0459	-3.4533	0.0080	0.0481
SKCM (tumor)	0.3898	0.4906	0.7070	0.3425	0.5525	0.6853
STAD	-0.2521	0.5910	0.7093	-0.2518	0.5939	0.6853
TGCT	-0.5761	0.0062	0.0459	-0.5721	0.0070	0.0481
THCA	0.2498	0.6815	0.7434	0.2933	0.6282	0.6853
SKCM (metastatic)	-0.4055	0.1502	0.4505	-0.3902	0.1694	0.5081

https://doi.org/10.1371/journal.pone.0237006.t004

have been identified, most of these age-related signatures were developed using either non-human tissues or small sample of tissues. Here, we utilized the gene expression data in the large GTEx database to identify common age-related genes as well as to construct a versatile across-tissue and tissue-specific transcriptional age calculator (RNAAgeCalc). We showed that transcriptional age acceleration was associated with lower mutation burden and lower mortality risk in TCGA cancer samples, and offered complementary information to DNAm age. Our results also indicated that racial difference was associated with transcriptional age. As majority of the samples in GTEx were Whites, future work included extending RNAAgeCalc to non-White samples.

RNAAgeCalc is available both as Bioconductor and Python packages as well as an interactive Shiny app. We anticipate that the calculator will be useful in the development of aging biomarker to advance our understanding on age-related diseases. These insights may ultimately inform development of novel treatments for age-related diseases.

Materials and methods

Implementation

Our software is capable of calculating tissue-specific transcriptional age for 26 types of tissues based on 8 choices of candidate feature sets described in this study. That is, users are able to

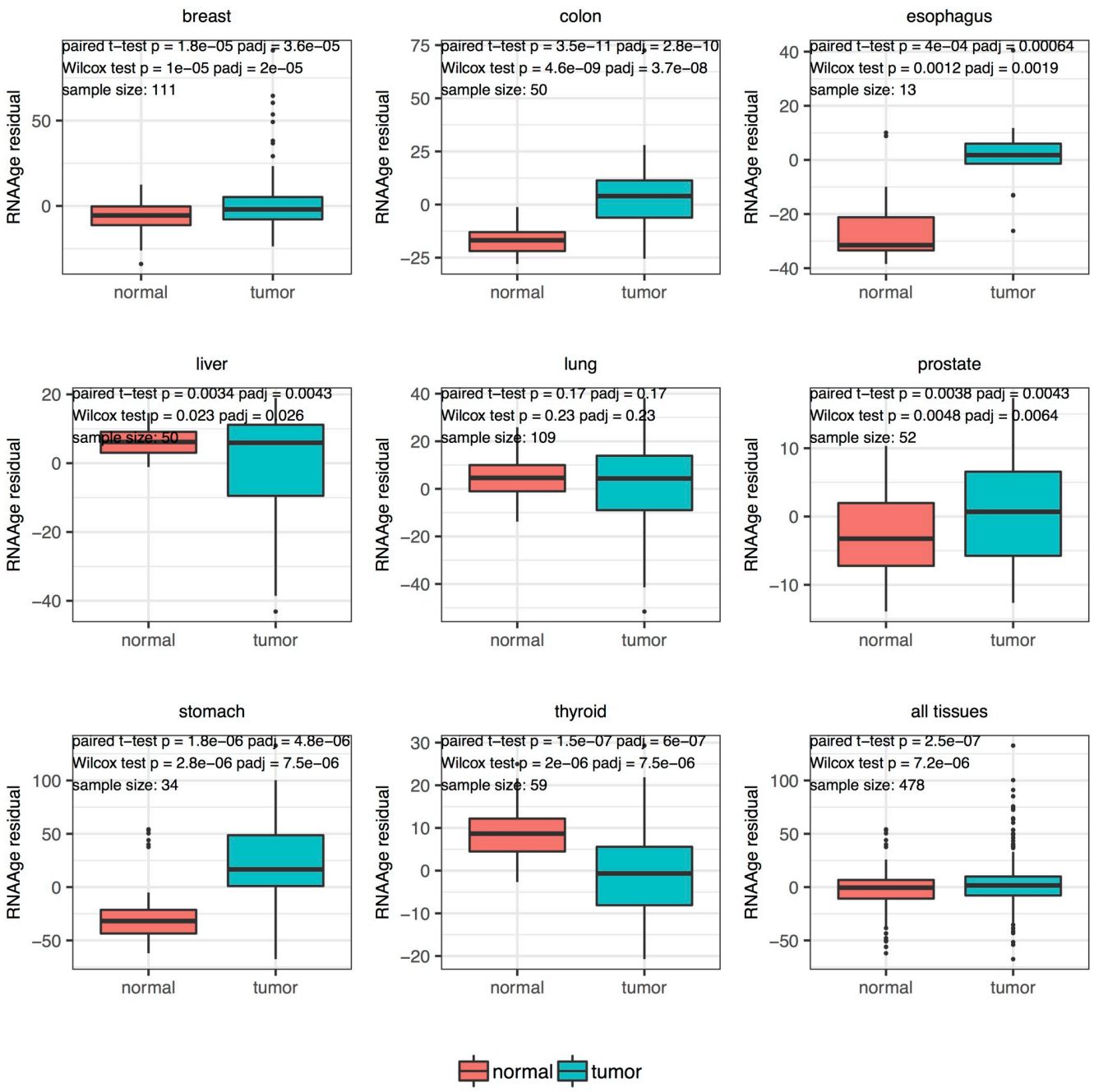

Fig 3. Age predictions on matched tumor and normal samples from TCGA.

https://doi.org/10.1371/journal.pone.0237006.g003

specify the choice of tissue, aging signatures, as well as sample type for racial consideration (i.e., whether to use calculators trained on all samples or Caucasian samples only). The software is implemented as follows. For each tissue, signature and sample type, we pre-trained the calculator using elastic net based on the GTEx samples. In the within-tissue age prediction section, we have demonstrated that elastic net model outperforms the other prediction algorithms. We saved the pre-trained model coefficients as internal data in the software. The software takes gene expression data as input and then match the input genes to the genes in the

internal data automatically. If the genes in user input data do not fully cover all the genes in the signature, imputation will be performed automatically by the `impute.knn()` function in Bioconductor package impute [37].

Our pre-trained models are computationally efficient because there is no need to re-train the elastic net model every time a new gene expression data is provided. In scenarios where the user input gene expression data covers the genes in the specified signature, transcriptional age computation only requires calculating the inner product of the coefficient vector and the gene expression. We take advantage of the vectorization in R and Python, which is computationally efficient. In scenarios where the user input gene expression data does not cover all the genes in the signature, the vast majority of computational time is in the imputation process. The time complexity of nearest neighbor imputation is $O(p\log p)$ for each gene, where p is the number of genes provided [37]. The computation could be expensive if the proportion of missing genes is large. In such scenarios, users should be cautious with the computed transcriptional age as it may be affected by the accuracy of imputation process.

GTEx data processing

To facilitate integrated analysis and direct comparison of multiple datasets, we utilized recount2 [38] version of GTEx data, where all samples were processed using the same analytical pipeline. FPKM values were calculated for each individual sample using `getRPKM()` function in Bioconductor package recount [38]. The benefit of using FPKM instead of raw RNASeq count to build up prediction model was that FPKM had been normalized for the total count and gene length, therefore enabling comparison across different RNA-Seq samples. The recount2 version of GTEx data contained 58,037 genes while the dermal fibroblast data [19] described in the Introduction section contained 27,142 genes. We studied genes which were measured on both recount2 version of GTEx data and dermal fibroblast data. Genes in recount2 were annotated using Ensembl ID whereas genes in dermal fibroblast data were annotated using RefSeq. We mapped Ensembl ID to RefSeq using Bioconductor package org.Hs.eg.db [39] (version 3.7.0) and only genes with one-to-one map were considered in the analysis, resulting in a total of 24,989 genes.

Meta-analysis to identify common gene signature

Within each tissue, RNASeq count data was imported from recount2 version of GTEx database and the gene ID subsetting and processing were exactly the same as discussed above. Tissues with fewer than 50 samples were omitted from the analysis as small sample size may lead to conservative and biased results. To avoid the influence of low count genes on the analysis result, genes with more than 20% samples having count per million (CPM) less than one were filtered out. Differential expression analyses with respect to chronological age were performed on each tissue using DESeq2 [40] based on raw gene counts, adjusting for gender (except for ovary, prostate, testis, uterus, and vagina tissues) and race (White vs non-White). In this paper, differentially expressed genes referred to the genes which were significantly associated with chronological age. To determine these genes, we performed the Wald test implemented in DESeq2 and obtained a p-value for each gene. The p-values of differential expression were further adjusted using the false discovery rate (FDR) procedure [29].

We adapted the binomial test of de Magalhaes et al. [12] to identify common age-related genes across tissues. Genes with FDR < 0.05 were considered significantly associated with age. For each individual gene, binomial test was performed with the p-values calculated by the

cumulative distribution function:

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \quad (1)$$

where n denotes the total number of tissues, k denotes the number of tissues in which the gene was positively (negatively) associated. The parameter p was estimated by the average proportion of positively (negatively) associated genes across tissues, resulting in value 13.86% (13.77%). The raw p-values in binomial tests were adjusted using FDR procedure, and statistically significant genes were identified at FDR < 0.05.

Enrichment analysis of age-related genes

To identify enriched pathways among our 1,616 age-related genes, 297 KEGG pathways [41] and 5,784 Gene Ontologies [42] (minimum and maximum number of genes for each gene set were 15 and 500, respectively) were tested using clusterProfiler software [43]. Hypergeometric tests were performed based on the 828 positively associated genes and 788 negative associated genes, respectively. The p-values from hypergeometric tests were adjusted using FDR procedure and gene sets with FDR < 0.05 were considered significant.

Transcriptional age prediction via machine learning models

For each individual tissue, we considered several machine learning models including elastic net [9], generalized boosted regression models (GBM) [24], random forest [25], support vector regression (SVR) with radial kernel [26] and ensemble LDA [19]. Most of these machine learning models have been implemented in R packages e.g., elastic net (R package glmnet [44]), generalized boosted regression (R package gbm [45], random forest (R package randomForest [46], support vector machine (R package e1071 [47]. For ensemble LDA, we adapted the python scripts provided by the authors [19] on Github. In all these models, chronological age was the response variable whereas the logarithm transformed FPKM were the predictors. The samples were first randomly split into 50%-50% training-test set, where the prediction algorithms were constructed on the training data and evaluated on the test data. The optimal parameters, namely alpha and lambda in elastic net, number of trees in GBM, cost and gamma in SVR, and bin size in ensemble LDA were selected by 10 fold cross validation in the training set. This 50%-50% training-test split and model evaluation were repeated 10 times. We considered the following candidate gene sets in constructing the prediction models. For each candidate feature set, we first took the subset of FPKM data corresponding to the pre-defined candidate genes. We then trained each machine learning model described above on the training data. The trained predictors were applied on the test subset to evaluate performance accuracy.

(i). Differentially expressed genes by DESeq2 [40] (denoted by DESeq2). Before training the prediction models, differential expression analysis on age was first performed on the training data. The gene filtering criterion and variables adjusted in differential expression analysis were the same as described in the above section. Instead of using all genes, only the most significant genes from differential expression analysis were used to train the prediction models. Here, we used top K differentially expressed genes ranked by the p-values from differential expression analysis. To investigate the influence of the number of genes on prediction accuracy, we considered within-tissue prediction using top 500, 1000, 1,500, 2,000 genes and compared their performances. S3 Appendix showed the prediction accuracy (Pearson/Spearman correlation between predicted

transcriptional age and chronological age, median/mean error) versus the number of top significant genes in prediction model. For some tissues the number of top differentially expressed genes did not have a huge impact on prediction accuracy, whereas for other tissues the prediction accuracy increased with the number of top significant genes. To reduce the computation cost, we considered the top 1,000 genes.

(ii). Genes highly correlated with chronological age by Pearson correlation (denoted by Pearson). Before training the prediction models, Pearson correlations between the logarithm transformed FPKM and chronological age were calculated on the training set. The correlation coefficients were then sorted by its absolute value in decreasing order, and only the top correlated (either positive or negative) genes were used to train the prediction models. Similar to (i), we also considered top 1,000 correlated genes and a comparison of the number of genes in the model was given in S3 Appendix.

(iii). Genes have large variance in expression across samples (denoted by Deviance). We adapted the gene selection strategy discussed in [19], in which a gene had at least a t_1-fold difference in expression between any two samples in the training set and at least one sample had expression level $>t_2$ FPKM to be included in the prediction models. t_1 and t_2 (typically 5 or 10) were the thresholds to control the degree of deviance of the genes. In our analysis, we used $t_1 = t_2 = 10$ for most tissues. For some tissues with large sample size, in order to maximize the prediction accuracy while maintaining low computation cost, we increased t_1 and t_2 such that the number of genes retained in the model was between 2,000 and 7,000. An alternative way of selecting genes with high variability is based on the coefficient of variation (denoted by CVar), which is defined as the ratio of standard deviation to mean. Since genes with higher counts could have larger variance compared to genes with lower counts (i.e., the well-known mean variance relationship in RNA-Seq data), selecting genes based on CVar could potentially reduce the bias toward high count genes. We compared these two approaches by performing within tissue transcriptional age prediction (S17 Table). To ensure a fair comparison, the number of top genes ranked by CVar was fixed to be the same as the number of genes ranked by Deviance. The prediction performance of both methods was comparable. Thus, we used Deviance for selecting large variation genes throughout this paper to be consistent with the criterion used in [19].

(iv). The 1,497 age-related genes of [18] (denoted by Peters).

(v). All genes after filtering out low count genes. Specifically, genes with more than 20% samples having CPM less than one were filtered out.

(vi). The 1,616 common age-related genes discussed in the meta-analysis section (denoted by GTExAge).

(vii). The 73 common age-related genes of de Magalhaes et al. (denoted by de Magalhaes) [12].

(viii). The 307 common age-related genes from the Ageing Gene Database (denoted by GenAge) [13].

The models were evaluated by the Pearson and Spearman correlation between predicted age and chronological age, median absolute error (median error) and mean absolute error (mean error) on the test samples, averaging over the 10 repetitions.

Associations with prior aging candidate genes

For the tissue-specific signatures, we extracted the genes reported in the original references and took their intersection with the genes available in GTEx. We investigated whether these prior aging candidate genes were significant in our DESeq2 analysis result (S10 Table). We then compared the sign of fold change of these prior genes to the sign in our DESeq2 result (S11 Table). Fisher exact tests were performed, which showed the signs are highly consistent ($p < 0.05$). The p-values of these prior genes in our DESeq2 analysis result is given in S3 Fig.

For each prior across-tissue candidate gene set, we first computed the p-value of each gene using the binomial test (see the meta-analysis section) as a summary measure of evidence the gene held as candidate common age-related gene. We then enumerated the proportion of genes which attained $p < 0.05$ within each prior across-tissue candidate gene set.

Comparisons of DNAm age versus transcriptional age on TCGA dataset

Illumina Human Methylation 450K annotation data were imported from the Broad GDAC Firehose and the DNAm age were obtained by analyzing the beta value using DNAm age calculator [7, 8, 10]. The TCGA RNASeq data was downloaded and processed from recount2 [38], following the same pipeline as described in the GTEx data processing section. Transcriptional age was obtained by applying the tissue-specific predictors based on all genes, DESeq2 and GTExAge candidate features on the corresponding tissue in TCGA. For skin cutaneous melanoma (SKCM), the tumor and metastatic samples were analyzed separately. For breast invasive carcinoma (BRCA), only female samples were analyzed. Age acceleration residual was defined as residual from regressing transcriptional age (or DNAm age) on chronological age. The significance of the correlation between age acceleration residual and mutation burden was evaluated by correlation tests. Cox proportional hazards model was fitted on the age acceleration residual and Wald test was performed on the estimated coefficient. Linear regression model was applied to compare age acceleration residual to cancer stage (ordinal covariate) and t-test was performed on the estimated coefficient. Paired sample t-test and Wilcoxon test was used to compare the transcriptional age from matched tumor and normal samples. The FDR adjusted p-values < 0.05 were considered statistically significant.

Supporting information

S1 Table. Summary of GTEx dataset.
(PDF)

S2 Table. Summary of differential expression analysis on each tissue.
(PDF)

S3 Table. List of GTExAge genes and pathway analysis results.
(XLSX)

S4 Table. Summary of candidate feature sets.
(PDF)

S5 Table. Within-tissue prediction results.
(XLSX)

S6 Table. Comparison of elastic net and ensemble LDA prediction accuracy.
(XLSX)

S7 Table. Across-tissue prediction results.
(XLSX)

S8 Table. Within-tissue prediction based on White samples only.
(XLSX)

S9 Table. Across-tissue prediction based on White samples only.
(XLSX)

S10 Table. Overlap between tissue-specific genes in GTEx and prior aging candidate genes.
(PDF)

S11 Table. Comparison of the sign of fold genes.
(PDF)

S12 Table. Comparison of GTEx aging genes versus prior candidate genes on within-tissue prediction.
(PDF)

S13 Table. Overlap between GTExAge genes and prior aging candidate genes.
(PDF)

S14 Table. Correlation between age acceleration residual and mutation burden (based on DESeq2 and GTExAge genes).
(PDF)

S15 Table. Coefficient and p-value of age acceleration residual from Cox regression (based on DESeq2 and GTExAge genes).
(PDF)

S16 Table. Coefficient and p-value of age acceleration residual versus stage from linear model (based on DESeq2 and GTExAge genes).
(PDF)

S17 Table. Comparison of Deviance and coefficient of variation in selecting genes with high variability.
(CSV)

S1 Fig. Heat-maps of Pearson correlation matrix between predicted age and chronological age (based on all genes and DESeq2 genes).
(PDF)

S2 Fig. Comparison of the prediction on White and non-White samples (based on GTEx-Age genes).
(PDF)

S3 Fig. Histogram of p-values of prior candidate genes in GTEx data computed from DESeq2.
(PDF)

S4 Fig. Ratio of transcriptional age to chronological age on matched tumor and normal samples from TCGA.
(PDF)

S1 Appendix. Comparison of elastic net and ensemble LDA.
(PDF)

S2 Appendix. Age predictions on matched tumor and normal samples by DNAm age calculator and related genes.
(PDF)

S3 Appendix. The accuracy of prediction algorithms using different number of genes.
(PDF)

Acknowledgments

The authors would like to thank Drs. Shannon Ellis and Jeffery Leek for sharing the GTEx metadata. The GTEx chronological age and self-reported race were protected data and the authors had been granted permission to use this information for research via dbGaP (https://www.ncbi.nlm.nih.gov/gap/).

Author Contributions

Conceptualization: Pei Fen Kuan.

Formal analysis: Xu Ren, Pei Fen Kuan.

Methodology: Xu Ren, Pei Fen Kuan.

Software: Xu Ren, Pei Fen Kuan.

Supervision: Pei Fen Kuan.

Writing – original draft: Xu Ren, Pei Fen Kuan.

References

1. Niccoli T, Partridge L. Ageing as a risk factor for disease. Current biology. 2012; 22(17):R741–R752. https://doi.org/10.1016/j.cub.2012.07.024 PMID: 22975005
2. Rodríguez-Rodero S, Fernández-Morera JL, Menéndez-Torre E, Calvanese V, Fernández AF, Fraga MF. Aging genetics and aging. Aging and disease. 2011; 2(3):186. PMID: 22396873
3. Pilling LC, Kuo CL, Sicinski K, Tamosauskaite J, Kuchel GA, Harries LW, et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. Aging (Albany NY). 2017; 9(12):2504. https://doi.org/10.18632/aging.101334
4. Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, Kumari M, et al. A genome-wide association study of aging. Neurobiology of aging. 2011; 32(11):2109–e15. https://doi.org/10.1016/j.neurobiolaging.2011.05.026 PMID: 21782286
5. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, et al. LDL-cholesterol concentrations: a genome-wide association study. The Lancet. 2008; 371(9611):483–491. https://doi.org/10.1016/S0140-6736(08)60208-1
6. Jansen I, Savage J, Watanabe K, Bryois J, Williams D, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer?s disease risk. Nature Genetics. 2019;. https://doi.org/10.1038/s41588-018-0333-3
7. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013; 14(10):3156. https://doi.org/10.1186/gb-2013-14-10-r115
8. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell. 2013; 49(2):359–367. https://doi.org/10.1016/j.molcel.2012.10.016 PMID: 23177740
9. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
10. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018; 10(4):573. https://doi.org/10.18632/aging.101414

11. Lu A, Quach A, Wilson J, Reiner A, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;.

12. De Magalhães JP, Curado J, Church GM. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics. 2009; 25(7):875–881. https://doi.org/10.1093/bioinformatics/btp073 PMID: 19189975

13. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. Human Ageing Genomic Resources: new and updated databases. Nucleic acids research. 2017; 46(D1):D1083–D1090. https://doi.org/10.1093/nar/gkx1042

14. Welle S, Brooks AI, Delehanty JM, Needler N, Thornton CA. Gene expression profile of aging in human muscle. Physiological genomics. 2003; 14(2):149–159. https://doi.org/10.1152/physiolgenomics.00049.2003 PMID: 12783983

15. Rodwell GE, Sonu R, Zahn JM, Lund J, Wilhelmy J, Wang L, et al. A transcriptional profile of aging in the human kidney. PLoS biology. 2004; 2(12):e427. https://doi.org/10.1371/journal.pbio.0020427 PMID: 15562319

16. Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, et al. Gene regulation and DNA damage in the ageing human brain. Nature. 2004; 429(6994):883. https://doi.org/10.1038/nature02661 PMID: 15190254

17. Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome biology. 2013; 14(7):R75. https://doi.org/10.1186/gb-2013-14-7-r75 PMID: 23889843

18. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. Nature communications. 2015; 6:8570. https://doi.org/10.1038/ncomms9570 PMID: 26490707

19. Fleischer JG, Schulte R, Tsai HH, Tyagi S, Ibarra A, Shokhirev MN, et al. Predicting age from the transcriptome of human dermal fibroblasts. Genome biology. 2018; 19(1):221. https://doi.org/10.1186/s13059-018-1599-6 PMID: 30567591

20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nature genetics. 2013; 45(6):580. https://doi.org/10.1038/ng.2653

21. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research. 2017; 45(W1):W98–W102. https://doi.org/10.1093/nar/gkx247 PMID: 28407145

22. Weindruch R, Kayo T, Lee CK, Prolla TA. Gene expression profiling of aging using DNA microarrays. Mechanisms of ageing and development. 2002; 123(2-3):177–193. https://doi.org/10.1016/S0047-6374(01)00344-X PMID: 11718811

23. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences. 2004; 101(25):9309–9314. https://doi.org/10.1073/pnas.0401994101

24. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001; p. 1189–1232. https://doi.org/10.1214/aos/1013203451

25. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

26. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–297. https://doi.org/10.1023/A:1022627411411

27. Yu Y, Xu T, Yu Y, Hao P, Li X. Association of tissue lineage and gene expression: conservatively and differentially expressed genes define common and special functions of tissues. BMC bioinformatics. 2010; 11(11):S1. https://doi.org/10.1186/1471-2105-11-S11-S1 PMID: 21172044

28. Gu P, Xu A. Interplay between adipose tissue and blood vessels in obesity and vascular dysfunction. Reviews in Endocrine and Metabolic Disorders. 2013; 14(1):49–58. https://doi.org/10.1007/s11154-012-9230-8 PMID: 23283583

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

30. Lin Q, Wagner W. Epigenetic aging signatures are coherently modified in cancer. PLoS genetics. 2015; 11(6):e1005334. https://doi.org/10.1371/journal.pgen.1005334 PMID: 26110659

31. Wagner W, Weidner CI, Lin Q. Do age-associated DNA methylation changes increase the risk of malignant transformation? Bioessays. 2015; 37(1):20–24. https://doi.org/10.1002/bies.201400063 PMID: 25303747

32. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics & chromatin. 2018; 11(1):37. https://doi.org/10.1186/s13072-018-0205-1

33. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome biology. 2014; 15(2):R37. https://doi.org/10.1186/gb-2014-15-2-r37 PMID: 24555846

34. Wang Z, Yin J, Zhou W, Bai J, Xie Y, Xu K, et al. Complex impact of DNA methylation on transcriptional dysregulation across 22 human cancer types. Nucleic Acids Research. 2020; 48(5):2287–2302. https://doi.org/10.1093/nar/gkaa041 PMID: 32002550

35. Dai JY, Wang X, Wang B, Sun W, Jordahl KM, Kolb S, et al. DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs. PLoS Genetics. 2020; 16(3):e1008667. https://doi.org/10.1371/journal.pgen.1008667 PMID: 32226005

36. Lim YC, Li J, Ni Y, Liang Q, Zhang J, Yeo GS, et al. A complex association between DNA methylation and gene expression in human placenta at first and third trimesters. PloS one. 2017; 12(7):e0181155. https://doi.org/10.1371/journal.pone.0181155 PMID: 28704530

37. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–525. https://doi.org/10.1093/bioinformatics/17.6.520 PMID: 11395428

38. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nature biotechnology. 2017; 35(4):319. https://doi.org/10.1038/nbt.3838 PMID: 28398307

39. Carlson M. org.Hs.eg.db: Genome wide annotation for Human; 2018.

40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014; 15(12):550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25. https://doi.org/10.1038/75556 PMID: 10802651

43. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012; 16(5):284–287. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

44. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010; 33(1):1. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

45. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models; 2019. Available from: https://CRAN.R-project.org/package=gbm.

46. Liaw A, Wiener M, et al. Classification and regression by randomForest. R news. 2002; 2(3):18–22.

47. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien; 2017. Available from: https://CRAN.R-project.org/package=e1071.