EpigeneticAgePipeline: an R package for comprehensive assessment of epigenetic age metrics from methylation microarrays

Stanislav Rayevskiy[1], Quinn Le[1], Julia Nguyen[1], Steven Q. Chen[1], Christina A. Castellani[1,2,3,4]*

[1] Department of Pathology and Laboratory Medicine, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada
[2] Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada
[3] McKusick-Nathans Institute, Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America
[4] Children's Health Research Institute, Lawson Research Institute, London, Ontario, Canada
* Corresponding author
Email: christina.castellani@schulich.uwo.ca

Abstract
Summary

Epigenetic age is a biological age estimation that is based off nuclear DNA methylation patterns. As an individual ages, methylation sites change in a predictable manner, allowing for the generation of epigenetic age clocks. EpigeneticAgePipeline is an R package that aims to provide a straightforward, out-of-the-box solution for measuring epigenetic age estimates (Horvath, skinHorvath, Hannum, PhenoAge/Levine, GrimAge), as well as two epigenetic age acceleration measures (DunedinPACE, GrimAge Acceleration). Further, EpigeneticAgePipeline imputes cell counts, as well as performs statistical analyses including principal component analysis, normalization, and generation of linear model-based residuals. The pipeline supports Illumina Infinium methylation microarray formats, including HumanMethylation27, HumanMethylation450, and HumanMethylationEPIC. Our package offers an all-in-one solution to generate epigenetic age estimates from input IDAT files or preprocessed methylation beta values.

Availability:

EpigeneticAgePipeline is implemented in R (3.5 or greater) and available on GitHub at https://github.com/CastellaniLab/EpigeneticAgePipeline

Supplementary Information:

PDF file submitted along with manuscript.

Keywords:
Epigenetics, Bioinformatics, DNA Methylation, Epigenetic Age

Introduction

The field of epigenomics has allowed for the understanding that environmental factors can significantly influence our genome and how particular regions of our genome are regulated and expressed. As the field has expanded, research into the association of epigenetic features, human health, disease risk, as well as aging has emerged, leading to the concept of epigenetic age and the creation of epigenetic age clocks. Epigenetic clocks are computational tools that estimate biological age by analyzing specific DNA methylation patterns that are associated with cellular aging across the genome. These clocks use measures of DNA methylation, chemical modifications that regulate gene expression without altering the underlying DNA sequence, to provide a look into the aging process. This is important because epigenetic age may be a better measure of an individual's health status than chronological age alone. *EpigeneticAgePipeline* aims to build off the functionality of other well established R packages such as dnaMethyAge [1] and methylclock [2] to provide additional functionality and significantly improve usability.

What makes *EpigeneticAgePipeline* unique is that it integrates the entire workflow, including data pre-processing, epigenetic age calculations, and downstream visualization tasks/final analysis into one seamless pipeline. The package allows for direct input of raw IDAT files, which are then processed via removal of cross-reactive probes, removal of sex probes, normalization, and sample quality control. Alternatively, users can input their own preprocessed methylation beta values. The package also allows for residual generation to account for cell count data and batch effects that arise during methylation data collection. Residual generation also includes the option to regress out principal components generated from the methylation dataset to account for additional sources of variation in the data.

Furthermore, *EpigeneticAgePipeline* provides a range of visualizations, including a variety of correlation and distribution plots that enhance the data's interpretability. These features, combined with the foundational functionalities of other well-established R packages, allow *EpigeneticAgePipeline* to fill a unique niche in computational epigenetics.

Features

EpigeneticAgePipeline generates five common epigenetic age clocks, which includes Horvath clock [3], Skin Horvath clock [4], Hannum clock [3], PhenoAge clock [3] and GrimAge [5] (Principal component version of GrimAge). Each one analyzes the methylation pattern of a unique set of CpG sites, which allows for more specific estimations depending on the tissue type and/or phenotype. Along with epigenetic age estimation, the package allows for the generation of two age acceleration estimates, DunedinPace [6] and GrimAge age acceleration [5].

The original Horvath clock was one of the first epigenetic clocks created and remains one of the "gold standard" epigenetic age measures [3]. The clock uses a set of 353 CpG sites and is designed to be used on a wide range of tissue types. The second original epigenetic clock is the Hannum clock [3]. This clock relies on a set of 71 CpG sites and is primarily designed for use on blood samples. Both the Horvath and Hannum clocks have been shown to be significant predictors of age- and age-related conditions [3]. After the creation of the Horvath clock,

the Skin Horvath clock was developed using a set of 391 CpG sites for better estimations from skin, blood, saliva, human fibroblasts, keratinocytes, buccal cell, endothelial cell and lymphoblastoid cell inputs. [4]. Another epigenetic age clock included is the Levine/PhenoAge clock which was created as a potential outperformer of other clocks concerning predictions of age-based phenotypic outcomes, it uses a set of 513 CpG sites [3].

GrimAge and GrimAge acceleration are measures primarily designed for blood samples and are based on a set of 78,464 CpG sites. GrimAge has been known to predict a high Mortality Hazard Ratio compared to other common epigenetic age measures [5]. The DunedinPACE measure is a newer measure of age acceleration, which has been known to show strong associations with phenotypic outcomes like morbidity, disability, and mortality [6]. DunedinPace was developed for use in blood samples and uses a set of 173 CpG sites [6].

An important step in the processing of raw methylation data from IDAT files is the removal of cross-reactive probes. For each of the 3 microarray formats accepted (27K, 450K and EPIC), a separate list of previously identified cross reactive probes are removed during this process [7,8,9].

EpigeneticAgePipeline also incorporates residual generation to allow for correction of important confounding variables. Principal component analysis is performed on beta values, and residuals are generated adjusting for the first 5 PCs (default) generated on the methylation data. Samples that are beyond 3 standard deviations from the mean in any of the first PCs are marked as outliers. Residual generation also incorporates adjustments for technical effects that commonly arise from methylation array measurements. The user may provide data for array variables such as Row, Column, Slide, and Batch. Depending on which are given, the model will dynamically construct a suitable formula for the regression (see Figure 1). There is also a built-in option to compute and adjust for cell count proportion estimates based off methylation data measured using EPIC or 450K microarray technology, this functionality uses the Houseman cell type imputation method [10]. Residuals generated when chronological age is set as a predictor for biological age effectively gives a measure of age acceleration.

A series of plots and visualizations are also included in the final output to present the correlations between epigenetic age measures and chronological age as well as the distribution of epigenetic age estimates for every sample. Raw data is exported as .txt file which can be used for further analysis.

Usage

The package's *main* function is the primary entry point for epigenetic age analysis. The user provides the directory containing methylation and covariate data, if available (age, sex, smoking status, batch effects, etc), along with selections concerning normalization of beta values before epigenetic age generation and the type of methylation array used.

The *generateResiduals* function acts as a tool for further analysis following epigenetic age generation. Specifically, the function will employ a linear model using user-specified variables and will output the corresponding residuals. Similar to the *main* function, the user provides the directory containing methylation and covariate data.

We have run our pipeline on a publicly available dataset, specifically GEO dataset GSE19764, which contains nuclear DNA methylation data for 23 samples measured using the Infinium MethylationEPIC array platform. The covariates included in these analyses were Sex, Age, Smoking status, and Array Row. Figures 2 and 3, as well as all *Supplementary Figures* were generated using our EpigeneticAgePipeline on the GSE19764 dataset.

Inputs: The package accepts methylation data either in the form of a .csv file containing methylation beta values for each respective sample, or a directory containing raw IDAT files.

Outputs: The package outputs raw epigenetic age/acceleration values (.txt and .md), raw residual results (.csv), age and epigenetic age correlation plots (.png, Figure 2), matrix correlation plots (.pdf, Figure 3), and an epigenetic age distribution plot (.png). Examples of all the plots can be found in the *Supplementary Information* files.

Specific information on installation procedures and file specificities, as well as vignettes are located within the GitHub readme located here: https://github.com/CastellaniLab/EpigeneticAgePipeline.

Conclusion

EpigeneticAgePipeline provides a comprehensive and user-friendly method to generate epigenetic age measures for those studying the effects of biological aging. Our pipeline better accounts for potential batch effects and confounding variables to minimize the effect of potential biases on the results in studies concerning biological aging. Our pipeline provides a complete workflow for epigenetic age estimation, starting with raw methylation data all the way to the generation of epigenetic age measures, incorporating a variety of quality control methods and visualizations.

EpigeneticAgePipeline serves as a valuable tool for investigating the relationships between epigenetic age measures and a wide range of health outcomes. Future updates will include the addition of emerging epigenetic aging measures/methods to ensure the pipeline remains relevant to this expanding field.

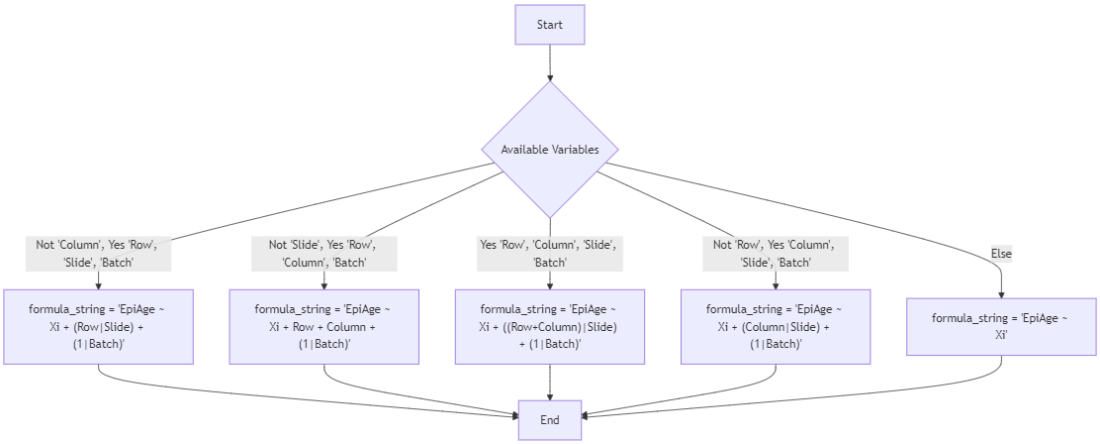

Figure 1: Flowchart illustrating formula construction for the linear model during residual generation (*generateResiduals* function). *Xi* represents all other user provided covariates. This schematic is based off presence of covariates "Row", "Column", "Slide", and "Batch".

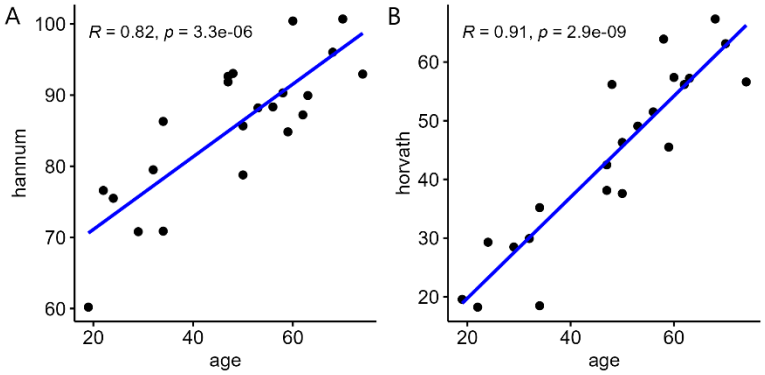

Figure 2: Scatter plots showing the relationship between Hannum (A) and Horvath (B) epigenetic age and chronological age for the GSE19764 dataset. The *main* function will output an analogous plot for each epigenetic age measurement if 'Age' is included as a covariate.

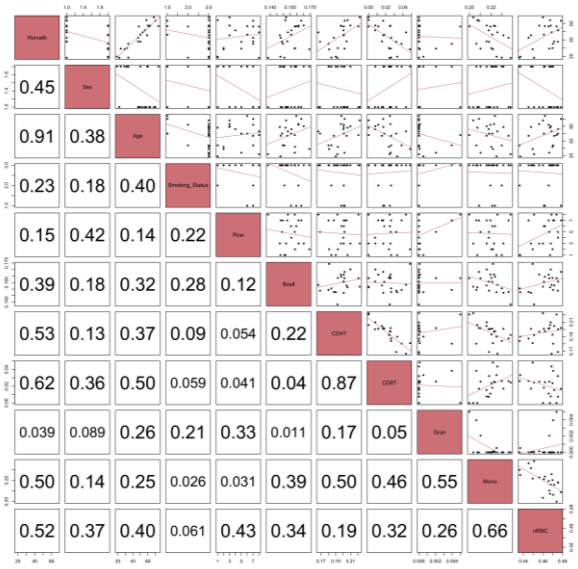

Figure 3: A correlation matrix generated for Horvath epigenetic age in the GSE19764 dataset. The boxes below the diagonal showcase the magnitude of the correlation between each respective pair of variables. The boxes above the diagonal showcase a scatterplot of pairwise variables with a linear trendline, giving the directionality and strength of the correlation. The *main* function will output an analogous plot for each measure of epigenetic age if covariates have been included.

References

[1] Y. Wang, O. Grant, X. Zhai, K. McDonald-Maier, and L. Schalkwyk, "Insights into ageing rates comparison across tissues from recalibrating cerebellum DNA methylation clock," *GeroScience*, vol. 46, Aug. 2023, doi: 10.1007/s11357-023-00871-w.

[2] D. Pelegí-Sisó, P. de Prado, J. Ronkainen, M. Bustamante, and J. R. González, "methylclock: a Bioconductor package to estimate DNA methylation age," *Bioinformatics*, vol. 37, no. 12, pp. 1759–1760, Jul. 2021, doi: 10.1093/bioinformatics/btaa825.

[3] D. V. Zavala *et al.*, "Epigenetic Age Acceleration and Chronological Age: Associations With Cognitive Performance in Daily Life," *J. Gerontol. Ser. A*, vol. 79, no. 1, p. glad242, Jan. 2024, doi: 10.1093/gerona/glad242.

[4] S. Horvath *et al.*, "Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and *ex vivo* studies," *Aging*, vol. 10, no. 7, pp. 1758–1775, Jul. 2018, doi: 10.18632/aging.101508.

[5] A. T. Higgins-Chen *et al.*, "A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking," *Nat. Aging*, vol. 2, no. 7, pp. 644–661, Jul. 2022, doi: 10.1038/s43587-022-00248-2.

[6] D. W. Belsky *et al.*, "DunedinPACE, a DNA methylation biomarker of the pace of aging," *eLife*, vol. 11, p. e73420, doi: 10.7554/eLife.73420.

[7] Y. Chen, S. Choufani, D. Grafodatskaya, D. T. Butcher, J. C. Ferreira, and R. Weksberg, "Cross-Reactive DNA Microarray Probes Lead to False Discovery of Autosomal Sex-Associated DNA Methylation," *Am. J. Hum. Genet.*, vol. 91, no. 4, pp. 762–764, Oct. 2012, doi: 10.1016/j.ajhg.2012.06.020.

[8] Y. Chen *et al.*, "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, Feb. 2013, doi: 10.4161/epi.23470.

[9] R. Pidsley *et al.*, "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling," *Genome Biol.*, vol. 17, p. 208, Oct. 2016, doi: 10.1186/s13059-016-1066-1.

[10] E. A. Houseman et al., "DNA methylation arrays as surrogate measures of cell mixture distribution," BMC Bioinformatics, vol. 13, no. 1, p. 86, May 2012, doi: 10.1186/1471-2105-13-86.