Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

methylclock: a Bioconductor package to estimate DNA methylation age

Dolors Pelegí-Sitó[1], Paula de Prado[1], Justiina Ronkainen[2], Mariona Bustamante[1,3], Juan R González[1,3,*]

[1] Barcelona Research Institute for Global Health (ISGlobal), Barcelona, Spain, [2]Center for Life Course Health Research, University of Oulu, Oulu, Finland, [3]CIBER in Epidemiology (CIBERESP), Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Ageing is a biological and psychosocial process related to diseases and mortality. It correlates with changes in DNA methylation (DNAm) in all human tissues. Therefore, epigenetic markers can be used to estimate biological age using DNA methylation profiling across tissues.
Results: We developed a Bioconductor package that allows computation of several existing DNAm adult/childhood and gestational age clocks. Functions to visualize the DNAm age prediction versus chronological age and the correlation between DNAm clocks are also available as well as other features such as missing data imputation of cell types estimates that are required for DNAm age clocks.
Availability: http://www.bioconducor.og and https://github.com/isglobal-brge/methylclock
Contact: juanr.gonzalez@isglobal.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ageing is a biological and psychosocial process related to diseases and mortality. It correlates with changes in DNA methylation (DNAm) in all human tissues. Therefore, epigenetic markers can be used to estimate biological age using DNA methylation profiling across tissues. Hannum[1] and Horvath[2] proposed the first epigenetic clocks to estimate DNAm age. Since their publication, several new epigenetic clocks have been proposed, for example clocks designed for specific tissues (cord blood, saliva, skeletal muscle, etc), or for specific ages, such as gestational age or childhood age. The deviation between age predicted from DNAm and chronological age has been proposed as a biomarker for ageing and has been related to overall survival and age-related diseases. Most of the existing DNAm age predictors are based on elastic net, which provides a set of cytosine-guanine (CpG) sites whose coefficients can be used to perform age prediction. Though its implementation is straightforward, there is no unified package that incorporates the wide range of estimators using into the same type of output. Some DNAm age calculators are based on online web-tools that make it difficult to integrate the computations in R/Bioconductor pipelines. Also, they require the methylation data to be in a specific format that is not optimal to manage the large amount of methylation markers available in the new methylation arrays. In order to overcome these limitations, we have developed a package, *methtylclock*, that uses R/Bioconductor infrastructures as the input as well as data frames or matrices importing data from other formats (.tsv, csv, …) for the researchers who are not familiar with those Bioconductor classes. The package also contains some functions to impute missing data since existing methods require complete cases. It also allows computation of the age-related functional decline of the immune system using the Extrinsic Epigenetic Age Acceleration (EEAA) estimator, and it has functions to estimate cell-type composition which is required to measure the age acceleration independently of age-related changes in the cellular composition of blood (Intrinsic Epigenetic Age Acceleration- IEAA). Functions to visualize the DNAm age prediction versus chronological age and the correlation between DNAm clocks are also available. All in all, *methylclock* encapsulates all the required methods to properly analyze the role of DNAm age in human traits.

1 Methods

DNAm clocks implemented in *methylclock* package are summarized in **Table 1**. The function *DNAmAge* computes childhood/adult age and biological age in years, *DNAmGA* estimates gestational age in months. **Supplementary Section 1** provides a more detailed information about the implemented clocks along with references. **Supplementary Section 2** pro-

vides a list of required packages. The package allows to have DNAm profiling in different formats including *ExpressionSet* (that allows to analyze data from GEO) *GenomicRatioSet* (used to encapsulate methylation data in *minfi* package) and data frame or tibbles for those researchers who want to import data from other formats.

Supplementary Section 3 provides information about other issues that are relevant to estimate DNAm age. These include data normalization, missing data and missing CpGs. Input data can be normalized using any of the existing standard methods such as QN, ASMN, PBC, SWAN, SQN, and BMIQ[3]. We have implemented BMIQ which is recommended by Horvath using a parallel version of this process using *BiocParallel* (Supplementary Section 3.1). This step is not mandatory, so that users can use normalized data and set the argument normalize to FALSE (default). DNAm clocks require complete cases. Our package imputes missing CpGs using *impute.knn* function. Large datasets can use median imputation by setting fastImp=TRUE. The required CpGs for each clock can be checked using *checkClocks* and *checkClocksGA* functions. By default, estimation of a given clock is performed when 80% of the CpGs are present, but this argument can be changed. Some biomarkers of age acceleration require estimation of white blood cell subtypes. This is implemented in *methylclock* using the *meffil* package and **Supplementary Section 3.5** describes the reference panels that are available for that purpose.

DNAmAge and *DNAmGA* return three biomarkers of age acceleration:

1) **ageAcc**: Difference between DNAm and chronological ages;
2) **ageAcc2:** Residuals obtained after regressing chronological age on DNAm age (similar to IEAA); and
3) **ageAcc3:** Residuals obtained after regressing chronological age on DNAm age adjusted for cell counts (similar to EEAA).

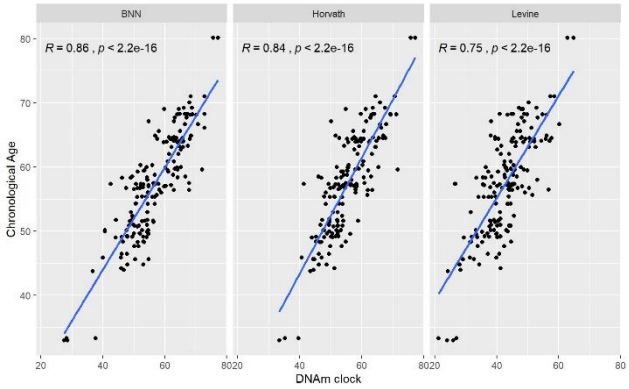
Finally, two functions can be used to inform users about the model's prediction performance. *plotDNAmAge* plots the regression plot between the chronological age and the one obtained by any clock joint with the R^2 value that can be used as a measure of agreement. *plotCorClocks* creates a panel with the same information of all estimated clocks with the correlation coefficient.

Table 1a. DNAm age estimates for gestational age (GA), childhood/adult age (AA) and biological age (BA) clocks available at *methylclock* package. References of each method can be seen in the Supplementary Material.

Type	Name	CpGs	Tissue	Array	Notes
AA	Horvath	353	Pan-tissue	27K and 450K	
AA	Hannum	71	Blood	27K and 450K	
AA	BNN	353	Pan-tissue	27K and 450K	Uses Horvath
AA	Horvath2	391	Skin+Blood	450K	
AA	PedBE	84	Buccal	450K	Age 0-20
GA	Knight	148	Cord Blood	27K and 450K	
GA	Bohlin	96	Cord Blood	450K	
GA	Mayne	62		27K and 450K	
GA	Lee				
	RPC	558	Placenta		Robust
	CPC	546	Placenta		Control
	RPC	396	Placenta		Complicated Gestations
BA	Levine	513	Blood	27-450K, EPIC	PhenoAge

3 Results

We have tested our package by analyzing several data sets from GEO repository (**Supplementary Sections 4 and 5).** Let us start by illustrating how to estimate adult age in healthy individuals. This is a good example



to evaluate how different clocks predict DNAm age. We downloaded data

Fig. 1. Correlation between DNAm and chronological age. The figure also includes the regression line and correlation coefficient of three method corresponds to GSE58045.

from GEO number GSE58045 which is a 27K experiment on 172 healthy individuals. **Figure 1** shows the correlation between BNN, Horvath and Levine estimators and the age reported in the GEO datasets. We can observe that both Horvath and BNN outperform Levine's clock, and that BNN slightly improves Horvath mainly because both methods are based on the same CpGs but BNN uses a Bayesian Network to predict age, which is a more sophisticated method than elastic net, since it can capture possible non-linear relationships. As a second example, we used a lung cancer study (GEO number GSE19711), in which 274 controls and 266 cases were analyzed with 27K array. The analysis revealed that age acceleration based on Horvath's clock was not associated with disease risk, while Levine's clock provided significant results for ageAcc (OR= 1.03, p= 0.005) and for ageAcc2 (OR= 1.05, p= 0.0003) while this association disappears when adjusted for cell counts(ageAcc3) (**Supplementary Section 4.4**).

4 Conclusion

methylclock is a new R/Bioconductor package for computing several DNAm clocks and plots that help to compare different estimators. The package can be integrated into pipelines designed to provide biological insights about ageing by using standard R/Bioconductor tools to perform association analyses.

Acknowledgments

We thank Janine Felix, Giulietta Monasso and Sylvain Severt for critical reading. This work has been partly supported by the Ministerio de Ciencia, Innovación y Universidades y Fondo Europeo de Desarrollo (RTI2018-100789-B-I00).

References

1. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367.
2. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, 3156.
3. Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J.C., and Chen, W. (2015). A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. Epigenetics *10*, 662–669.