# Manual for the New Online Epigenetic Age Calculator

**Steve Horvath (shorvath at mednet.ucla.edu)**

**Ake T. Lu (akekaikailu@gmail.com)**

This tutorial illustrates how to calculate DNA methylation age using the new version of the online epigenetic clock calculator.

The online clock can be found here:

https://dnamage.genetics.ucla.edu/new

Mandatory input: A (compressed) file with beta values, e.g. measured on the Illumina EPIC array or the 450K array or the 27k array. Optionally, you can compress the comma delimited file (.csv files) into a file that ends either with .zip or with .bz2. Other compression formats cannot yet be used.

You need to upload a file with beta values that contains ALL cg numbers listed in the file datMiniAnnotation3.csv.

Note that your file of beta values must contain at least 30084 rows corresponding to all the cg numbers in datMiniAnnotation3. Some of these CpGs are NOT present on the latest Illumina arrays. In this case simply create a row that contains the cg number and NA (missing values.

## The next steps are

(1) Go to https://dnamage.genetics.ucla.edu/home

(2) Select the panel "New Methylation Age Calculator". More details of

the instructions can be found in panel "Important Hints".

You need to upload a sample annotation file that specifies Age, Tissue, Female. Note the precise spelling of the variables

"**Age**" (starting with capital A), "**Female**" (with values 1 for female, 0 for male, NA for missing info), "**Tissue**".

*Make sure that the rows (samples) in the sample annotation file have the same order as the columns (samples) in the methylation file.* If you provide a sample annotation file then you will obtain the following variables:

- AgeAccelerationResidual=the recommended age acceleration measure based on a linear regression model.
- AgeAccelerationDiff=DNAmAge-Age

- predictedGender (based on the DNAm levels of X chromosomal markers)
- predictedTissue and probabilities that the sample comes from various tissues).

Advanced Analysis

If you used the 450K or the EPIC array platform you can get a host of additional output by selecting the AdvancedAnalysis option on the webpage. In this case, the software will output

- additional measures of biological age in blood: GrimAge, PhenoAge
- surrogate biomarkers of plasma proteins: PAI1, GDF15, CystatinC based on blood methylation
- estimates of smoking packyears based on blood methylation
- DNAm based estimator of telomere length
- estimates of blood cell counts
- estimated proportion of neurons based on brain methylation data
- different measures of epigenetic age acceleration.

Main citation for this software

> Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115 PMID: 24138928

## Legal Disclaimer

The epigenetic clock software is strictly a research tool. Our UCLA team has made every attempt to ensure the accuracy and reliability of the information provided by the epigenetic clock software. However, the information is provided "as is" without warranty of any kind. Neither UCLA nor the investigators accept any responsibility or liability for the accuracy, content, completeness, legality, or reliability of the information provided by this software.

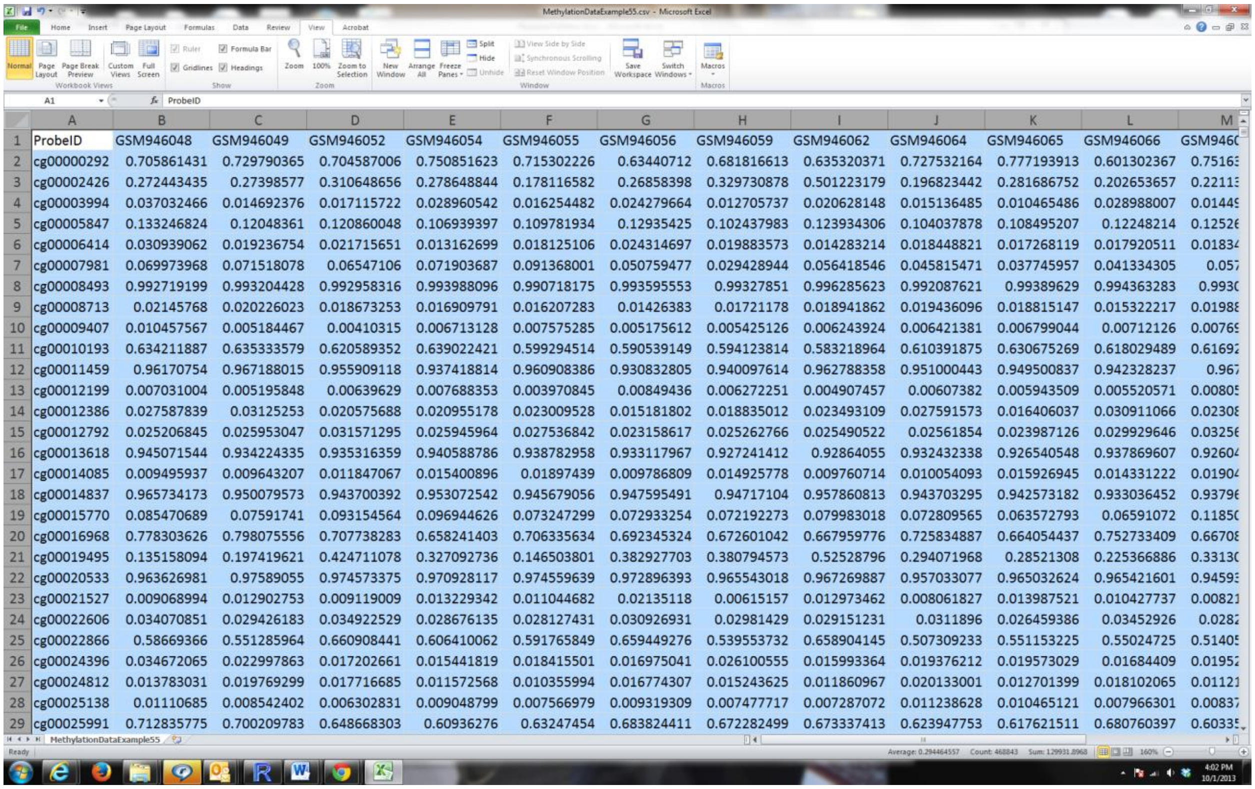
No warranties, promises and/or representations of any kind, expressed or implied, are given as to the nature, standard, accuracy or otherwise of the information provided by the software nor to the suitability or otherwise of the information to your particular circumstances.

# Contents

**Instructions**

Go to the webpage: https://dnamage.genetics.ucla.edu/new

The following screen shot shows that this input file is a comma delimited Excel file whose first column reports probe identifiers. The remaining columns correspond to samples (i.e. DNA meth arrays) for which DNAm age will be estimated.



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ProbeID | GSM946048 | GSM946049 | GSM946052 | GSM946054 | GSM946055 | GSM946056 | GSM946059 | GSM946062 | GSM946064 | GSM946065 | GSM946066 | GSM9460 |
| 2 | cg00000292 | 0.705861431 | 0.729790365 | 0.704587006 | 0.750851623 | 0.715302226 | 0.63440712 | 0.681816613 | 0.635320371 | 0.727532164 | 0.777193913 | 0.601302367 | 0.75163 |
| 3 | cg00002426 | 0.272443435 | 0.27398577 | 0.310648656 | 0.278648844 | 0.178116582 | 0.26858398 | 0.329730878 | 0.501223179 | 0.196823442 | 0.281686752 | 0.202653657 | 0.22113 |
| 4 | cg00003994 | 0.037032466 | 0.014692376 | 0.017115722 | 0.028960542 | 0.016254482 | 0.024279664 | 0.012705737 | 0.020628148 | 0.015136485 | 0.010465486 | 0.028988007 | 0.01449 |
| 5 | cg00005847 | 0.133246824 | 0.12048361 | 0.120860048 | 0.106939397 | 0.109781934 | 0.12935425 | 0.102437983 | 0.123934306 | 0.104037878 | 0.108495207 | 0.12248214 | 0.12526 |
| 6 | cg00006414 | 0.030939062 | 0.019236754 | 0.021715651 | 0.013162699 | 0.018125106 | 0.024314697 | 0.019883573 | 0.014283214 | 0.018448821 | 0.017268119 | 0.017920511 | 0.01834 |
| 7 | cg00007981 | 0.069973968 | 0.071518078 | 0.06547106 | 0.071903687 | 0.091368001 | 0.050759477 | 0.029428944 | 0.056418546 | 0.045815471 | 0.037745957 | 0.041334305 | 0.057 |
| 8 | cg00008493 | 0.992719199 | 0.993204428 | 0.992958316 | 0.993988096 | 0.990718175 | 0.993595553 | 0.99327851 | 0.996285623 | 0.992087621 | 0.99389629 | 0.994363283 | 0.9930 |
| 9 | cg00008713 | 0.02145768 | 0.020226023 | 0.018673253 | 0.016909791 | 0.016207283 | 0.01426383 | 0.01721178 | 0.018941862 | 0.019436096 | 0.018815147 | 0.015322217 | 0.01988 |
| 10 | cg00009407 | 0.010457567 | 0.005184467 | 0.00410315 | 0.006713128 | 0.007575285 | 0.005175612 | 0.005425126 | 0.006243924 | 0.006421381 | 0.006799044 | 0.00712126 | 0.00769 |
| 11 | cg00010193 | 0.634211887 | 0.635333579 | 0.620589352 | 0.639022421 | 0.599294514 | 0.590539149 | 0.594123814 | 0.583218964 | 0.610391875 | 0.630675269 | 0.618029489 | 0.61692 |
| 12 | cg00011459 | 0.96170754 | 0.967188015 | 0.955909118 | 0.937418814 | 0.960908386 | 0.930832805 | 0.940097614 | 0.962788358 | 0.951000443 | 0.949500837 | 0.942328237 | 0.967 |
| 13 | cg00012199 | 0.007031004 | 0.005195848 | 0.00639629 | 0.007688353 | 0.003970845 | 0.00849436 | 0.006272251 | 0.004907457 | 0.00607382 | 0.005943509 | 0.005520571 | 0.00805 |
| 14 | cg00012386 | 0.027587839 | 0.03125253 | 0.020575688 | 0.020955178 | 0.023009528 | 0.015181802 | 0.018835012 | 0.023493109 | 0.027591573 | 0.016406037 | 0.030911066 | 0.02308 |
| 15 | cg00012792 | 0.025206845 | 0.025953047 | 0.031571295 | 0.025945964 | 0.027536842 | 0.023158617 | 0.025262766 | 0.025490522 | 0.02561854 | 0.023987126 | 0.029929646 | 0.03256 |
| 16 | cg00013618 | 0.945071544 | 0.934224335 | 0.935316359 | 0.940588786 | 0.938782958 | 0.933117967 | 0.927241412 | 0.92864055 | 0.932432338 | 0.926540548 | 0.937869607 | 0.92604 |
| 17 | cg00014085 | 0.009495937 | 0.009643207 | 0.011847067 | 0.015400896 | 0.01897439 | 0.009786809 | 0.014925778 | 0.009760714 | 0.010054093 | 0.015926945 | 0.014331222 | 0.01904 |
| 18 | cg00014837 | 0.965734173 | 0.950079573 | 0.943700392 | 0.953072542 | 0.945679056 | 0.947595491 | 0.94717104 | 0.957860813 | 0.943703295 | 0.942573182 | 0.933036452 | 0.93796 |
| 19 | cg00015770 | 0.085470689 | 0.07591741 | 0.093154564 | 0.096944626 | 0.073247299 | 0.072933254 | 0.072192273 | 0.079983018 | 0.072809565 | 0.063572793 | 0.06591072 | 0.11850 |
| 20 | cg00016968 | 0.778303626 | 0.798075556 | 0.707738283 | 0.658241403 | 0.706335634 | 0.692345324 | 0.672601042 | 0.667959776 | 0.725834887 | 0.664054437 | 0.752733409 | 0.66708 |
| 21 | cg00019495 | 0.135158094 | 0.197419621 | 0.424711078 | 0.327092736 | 0.146503801 | 0.382927703 | 0.380794573 | 0.52528796 | 0.294071968 | 0.28521308 | 0.225366886 | 0.33130 |
| 22 | cg00020533 | 0.963626981 | 0.97589055 | 0.974573375 | 0.970928117 | 0.974559639 | 0.972896393 | 0.965543018 | 0.967269887 | 0.957033077 | 0.965032624 | 0.965421601 | 0.94593 |
| 23 | cg00021527 | 0.009068994 | 0.012902753 | 0.009119009 | 0.013229342 | 0.011044682 | 0.02135118 | 0.00615157 | 0.012973462 | 0.008061827 | 0.013987521 | 0.010427737 | 0.00821 |
| 24 | cg00022606 | 0.034070851 | 0.029426183 | 0.034922529 | 0.028676135 | 0.028127431 | 0.030926931 | 0.02981429 | 0.029151231 | 0.0311896 | 0.026459386 | 0.03452926 | 0.0282 |
| 25 | cg00022866 | 0.58669366 | 0.551285964 | 0.660908441 | 0.606410062 | 0.591765849 | 0.659449276 | 0.539553732 | 0.658904145 | 0.507309233 | 0.551153225 | 0.55024725 | 0.51405 |
| 26 | cg00024396 | 0.034672065 | 0.022997863 | 0.017202661 | 0.015441819 | 0.018415501 | 0.016975041 | 0.026100555 | 0.015993364 | 0.019376212 | 0.019573029 | 0.01684409 | 0.01952 |
| 27 | cg00024812 | 0.013783031 | 0.019769299 | 0.017716685 | 0.011572568 | 0.010355994 | 0.016774307 | 0.015243625 | 0.011860967 | 0.020133001 | 0.012701399 | 0.018102065 | 0.01121 |
| 28 | cg00025138 | 0.01110685 | 0.008542402 | 0.006302831 | 0.009048799 | 0.007566979 | 0.009319309 | 0.007477717 | 0.007287072 | 0.011238628 | 0.010465121 | 0.007966301 | 0.00837 |
| 29 | cg00025991 | 0.712835775 | 0.700209783 | 0.648668303 | 0.60936276 | 0.63247454 | 0.683824411 | 0.672282499 | 0.673337413 | 0.623947753 | 0.617621511 | 0.680760397 | 0.60335 |

## How to upload the data?

## Upload form

In the online form, enter your

1. Name:
2. Organization:
3. Email address. Explicitly enter it. Do not leave it empty. The results will be sent to this email address. Make sure it works.
4. Data file: Select the comma delimited file that contains your data. As mentioned before you can upload a zipped version of this file.

## Strategies for uploading very large data sets

Please take a note of the **upper limit** when it comes to uploading files. If you have a large data set that exceeds these limits then I recommend the strategies below. If you have a very large data set, start with strategy 2 and then move to strategy 1.

Strategy 1: Compress the file into a file that ends either with .zip or with .bz2. Other compression formats cannot yet be used.

Strategy 2: Turn your Illumina EPIC or 450K data into a "reduced" file that only contains probes that can be found in the file datMiniAnnotation3.csv (which is on our webpage). This does not result in any information loss since the epigenetic clock only uses probes that can be found in this file. After implementing this step, compress the resulting file (i.e. apply Strategy 1).

CpG probes that were not measured in your data set (e.g. are not present on the 450K array) should lead to a row filled with NAs.

Here is some relevant R code that assume your large data file is called "dat0" and the first column of dat0 contains the probe identifiers.

## R code for preparing the data

Below is code  for preparing data for submission to the online clock based on the file datMiniAnnotation3.

It reduced your data to roughly 30k CpGs in the file datMiniAnnotation3.

Steve's version:

```
setwd("PathToYourDirectory")
datMiniAnnotation=read.csv("datMiniAnnotation3.csv")

match1=match(datMiniAnnotation[,1],dat0[,1])
# the following should contain lots of integers…
match1[1:20]

dat0Short=dat0[match1,]
dat0Short[,1]=as.character(datMiniAnnotation[,1])

write.table(dat0Short,file="dat0NoobShortMini3.csv",row.names=FALSE,
sep=",")
```

Comment: it is a good idea to zip the .csv files before submission.

Ake's code

```
rm(list=ls())
options(stringsAsFactors=F)
set.seed('12345')
setwd("PathToYourDirectory")
#
#your input methyarray file
#
mycpg='mycpg.csv'
#
#your output filename after adding missing CpGs
#
out.csv='mycpg_updated.csv'
#
#PGM
#
input=read.csv(mycpg)
colnames(input)[1]='ProbeID'
ann=read.csv('datMiniAnnotation3.csv')
cpgs=input$ProbeID
check=is.element(ann$Name,cpgs)
table(check)
#
miss.cpg=ann$Name[!check]
#add NA
input.subject=colnames(input)[-1]
nsubject=dim(input)[2]-1
nmiss.cpg=length(miss.cpg)
#add subjects into the columns
add=data.frame(matrix(data=NA,nrow=nmiss.cpg,ncol=dim(input)[2]))
names(add)=names(input)
add[,1]=miss.cpg
#
#combine
#
output=rbind(input,add)
output=subset(output,ProbeID %in% ann$Name)
cat('check my new input dimension\n')
print(dim(output))
dim(ann)[1]==dim(output)[1]
#
head(output[,1:3])
output$permu=sample(dim(output)[1])
output=output[order(output$permu),]
output$permu<-NULL
head(output[,1:3])
#
write.table(output,out.csv,sep=',',row.names=F,quote=F)
```

Strategy 3: Split the data into batches, e.g. batches of 500 samples each. Next apply strategies 1 or 2.

Strategy 4: Email Steve Horvath or Yining Zhao to increase the upload limit for you.

## Normalization, imputation

We typically use *preprocessNoob* or *preprocessQuantile* from the minfi package but other methods work equally well. We recommend to avoid missing data at all costs, i.e. to ignore detection p values.

Additional buttons for the DNAm Age calculator allow you to check whether you want to normalize the data. It is strongly recommended to use the default setting (i.e. check "Normalize Data") since it often improves the predictive accuracy.

I have noticed that some users don't select this option since they think that they have their own superior normalization method. You should still check "Normalize Data". Reason: your normalization method has a different goal from my normalization method. The purpose of my normalization method is to make your data comparable to the training data of the epigenetic clock.

I advise against using the fast imputation method. However, if you have hundreds of samples with missing data and want to get a quick result then check "Fast Imputation".

## Uploading the sample annotation file

Sample annotation format
This sample annotation file is optional. Please upload it if you want to
a) obtain various measures of age acceleration,
b) allow the function to do some basic quality checks (e.g. check of gender, tissue).

Requirements: The sample annotation file should be comma delimited text file whose rows correspond to samples (e.g. human subjects). *Make sure that the rows (samples) in the sample annotation file have the same order as the columns (samples) in the methylation file.*

1) Not necessary but highly recommended: The first column should report the sample identifiers (matching those of the DNA methylation data, e.g. "Subject1", etc).
2) Mandatory: a column whose name is spelled **"Age"**. This column should report the (chronological) age in years, e.g. 0 for a newborn, 0.5 encodes a 6 month old child, 30 for a 30 year old. Prenatal samples would get a negative value, i.e. -.5 for a sample measured half a year before the expected birth. If you don't have age values, simply fill up the column with "NA".
3) Optional: I strongly recommend that you include gender information since this allows us to check whether the data are properly normalized etc. Toward this end, please insert a column called "**Female**" (note the capitalization) which takes a value of 1 if the subject is female, 0 if the subject is male, and NA if the information is not available. If you don't use ones or zeros, you will

get an error message. The calculator will output a column called "predictedGender". If the gender prediction does not match the known gender then there may be data quality issues.
4) Optional: I strongly recommend that you include a column that reports the DNA source (e.g. tissue). Toward this end, please insert a column called "**Tissue**" (note the capitalization) which takes a descriptive value. The tissue prediction tool is not yet published and its predictions should be interpreted with all due caution. I include this early version since it may help you identify mislabeled/suspicious samples.

Check whether one of the following descriptive terms matches your DNA source. If so, please use it. Otherwise simply report the best name that describes your DNA source.

```
[1] " Vasc.Endoth(Umbilical)"
 [2] "Ape WB"
 [3] "Blood CD4 Tcells"
 [4] "Blood CD4+CD14"
 [5] "Blood Cell Types"
 [6] "Blood Cord"
 [7] "Blood PBMC"
 [8] "Blood WB"
 [9] "Bone"
[10] "Brain Cerebellar"
[11] "Brain CRBLM"
[12] "Brain FCTX"
[13] "Brain Occipital Cortex"
[14] "Brain PONS"
[15] "Brain Prefr.CTX"
[16] "Brain TCTX"
[17] "Breast"
[18] "Breast NL"
[19] "Buccal"
[20] "Cartilage Knee"
[21] "Colon"
[22] "Dermal fibroblast"
[23] "Epidermis"
[24] "Fat Adip"
[25] "Gastric"
[26] "GlialCell"
[27] "Head+Neck"
[28] "Heart"
[29] "Kidney"
[30] "Liver"
[31] "Liver "
[32] "Lung"
[33] "MSC"   note that this stands for mesenchymal stromal cells
[34] "Muscle"
[35] "Neuron"
[36] "Placenta"
```

```
[37] "Prostate NL"
[38] "Saliva"
[39] "Sperm"
[40] "Stomach"
[41] "Thyroid"
[42] "Uterine Cervix"
[43] "Uterine Endomet"
```

The software will output a column called **predictedTissue**, which reports the predicted DNA source, i.e. one of the above mentioned DNA sources. Future versions of the age predictor will report more potential DNA sources.

## After you push the submit button

Push the "Submit" button. After a few minutes you will receive an email with the subject heading "Your Processing Result" that contains two attachments. The first attached file, whose name ends with "...output.csv" is a comma delimited file (which can be opened with Excel).

**How long does it take to get an email after your submitted your data?**

That depends on your sample size and whether or not you want the software to normalize the data. If you don't normalized the data, you should get an email within a couple of minutes. In contrast, normalizing several hundred samples could take several hours.

If you don't get any email, it means that your data crashed the R program. In this case, please carefully look at your input data. Do they meet the requirements? Maybe your methylation data set contains non-numeric variables (apart from the identifiers in the first column).

## OUTPUT of the clock with citations

If you measured Illumina 450K or EPIC data then we recommend that you select the advanced analysis option. Side note: If you have more than say 100 samples then use the data compression strategies 2 and 1 described in Strategies for uploading very large data sets.

The advanced option of the epigenetic clock software implements several epigenetic biomarker of lifespan and healthspan: pan tissue clock (Horvath 2013), blood based epigenetic clocks (Hannum 2013), skin and blood clock (Horvath 2018), estimators of phenotypic age (Levine 2017) and mortality risk DNAmGrimAge (Lu 2019a), estimator of telomere length: DNAmTL (Lu 2019b) , epigenetic estimators of plasma proteins e.g. DNAmPAI1, DNAm estimate of smoking pack years (DNAmPACKYRS).

**Variable naming convention:**

1) If a column name ends with "AdjAge" it means that it was adjusted for chronological age by forming a raw residual. Example. DNAmPAI1AdjAge=residuals(lm(DNAmPAI1~Age))
2) If a variable name starts with "AgeAccel" it means that we are dealing with an epigenetic measure of age acceleration, i.e. the variable was again adjusted for chronological age Example: AgeAccelerationResidual=age adjusted version of DNAmAge, i.e. AgeAccelerationResidual=residuals(lm(DNAmAge~Age))
   Note that

Main Reference for the online epigenetic clock software:

- Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115 PMID: 24138928

DNAmAge = pan tissue clock from Horvath 2013, AgeAccelerationResidual= measures of epigenetic age acceleration (residual).

- Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115 PMID: 24138928

DNAmGrimAge, DNAmPAI1, DNAmGDF15, DNAmPACKYRS, AgeAccelGrim etc.

DNAmADM DNAmB2M DNAmCystatinC DNAmGDF15

DNAmLeptin DNAmPACKYRS DNAmTIMP1

- Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019 Jan 21;11(2):303-327. doi: 10.18632/aging.101684. PMID: 30669119 PMCID: PMC6366976

Comment: DNAmGrimAge is *not* an age estimator. Chronological age and gender are part of its definition. Rather DNAmGrimAge estimates mortality risk (in units of years). DNAmGrimAge was trained in blood. Not sure whether it applies to other sources of DNA. I would be nervous about applying it to cell cultures.

The DNAmPAI1= epigenetic estimator of plasminogen activator inhibitor 1 in units of pg/ml.  But the units don't make sense. Just use it as an ordinal variable.

DNAmAgeSkinBloodClock= skin & blood clock estimate of age

- Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, Felton S, Matsuyama M, Lowe D, Kabacik S, Wilson JG, Reiner AP, Maierhofer A, Flunkert J, Aviv A, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Ferrucci L, Matsuyama S, Raj K. (2018) Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY). 2018 Jul 26;10(7):1758-1775. doi: 10.18632/aging.101508. PMID: 30048243 PMCID: PMC6075434

DNAmPhenoAge: phenotypic age estimate, AgeAccelPheno

- Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, Aviv A, Lohman K, Liu Y, Ferrucci L, Horvath S. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018 Apr 18;10(4):573-591. doi: 10.18632/aging.101414. PMID: 29676998 PMCID: PMC5940111

DNAmAgeHannum= blood based estimator of age from Hannum 2013, AgeAccelerationResidualHannum

- Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013 Jan 24;49(2):359-67.

EEAA, IEAA, BioAge4HAStatic: extrinsic epigenetic age acceleration, intrinsic epigenetic age acceleration

- Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, Ritz BR, Chen B, Lu AT, Rickabaugh TM, Jamieson BD, Sun D, Li S, Chen W, Quintana-Murci L, Fagny M, Kobor MS, Tsao PS, Reiner AP, Edlefsen KL, Absher D, Assimes TL (2016) An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016 Aug 11;17(1):171. doi: 10.1186/s13059-016-1030-0. PMID: 27511193
- Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, Roetker NS, Just AC, Demerath EW, Guan W, Bressler J, Fornage M, Studenski S, Vandiver AR, Moore AZ, Tanaka T, Kiel DP, Liang L, Vokonas P, Schwartz J, Lunetta KL, Murabito JM, Bandinelli S, Hernandez DG, Melzer D, Nalls M, Pilling LC, Price TR, Singleton AB, Gieger C, Holle R, Kretschmer A, Kronenberg F, Kunze S, Linseisen J, Meisinger C, Rathmann W, Waldenberger M, Visscher PM, Shah S, Wray NR, McRae AF, Franco OH,

Hofman A, Uitterlinden AG, Absher D, Assimes T, Levine ME, Lu AT, Tsao PS, Hou L, Manson JE, Carty CL, LaCroix AZ, Reiner AP, Spector TD, Feinberg AP, Levy D, Baccarelli A, van Meurs J, Bell JT, Peters A, Deary IJ, Pankow JS, Ferrucci L, Horvath S (2016) DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). 2016 Sep 28;8(9):1844-1865. doi: 10.18632/aging.101020. PMID: 27690265

BioAge4HAStatic

This age estimator is modified version of the predicted age measure based on the 71 CpGs in Hannum 2013. It only applies to BLOOD tissue. The calibration was defined as a linear transformation that ensures that the predicted age is aligned with chronological age. Message: this age estimator makes use of Age. Thus, the error between BioAge4HA and Age is reduced. Note that BioAge4HAStatic can only be calculated for those samples whose chronological age is available in the variable "Age". If Age is not available or all samples have the same age (zero variance) consider using DNAmAgeHannum or another measure (DNAmPhenoAge).

Why is BioAge4HAStatic important? Reason: it gives rise to the measure of extrinsic epigenetic age acceleration EEAA after adjustment for chronological age. EEAA is the raw residual that results from regressing BioAge4HAStatic on chronological age.

PlasmaBlast, CD8pCD28nCD45RAn, CD8.naive, CD4.naive=abundance estimates of special blood cell types e.g. CD8+CD28-CD45RA- T cells (i.e. exhausted T Cells) or naïve CD8+ T cells

- Horvath S, Levine AJ (2015) HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis. pii: jiv277. PMID: 25969563
- Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, Ritz BR, Chen B, Lu AT, Rickabaugh TM, Jamieson BD, Sun D, Li S, Chen W, Quintana-Murci L, Fagny M, Kobor MS, Tsao PS, Reiner AP, Edlefsen KL, Absher D, Assimes TL (2016) An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016 Aug 11;17(1):171. doi: 10.1186/s13059-016-1030-0. PMID: 27511193

CD8T, CD4T, NK, Bcell, Mono, Gran= proportions of blood cell types: monocytes, granulocytes etc

- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012, 13:86 doi:10.1186/1471-2105-13-86
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of

Infinium DNA methylation microarrays. Bioinformatics. 2014 May 15;30(10):1363-9. doi: 10.1093/bioinformatics/btu049.

### PlasmaBlastAdjAge, CD8pCD28nCD45RAnAdjAge, CD8.naiveAdjAge, CD4.naiveAdjAge

These are age adjusted versions of PlasmaBlast, CD8pCD28nCD45RAn, CD8.naive, CD4.naive. In other words, these are residuals resulting from a linear model that regresses the respective cell abundance measure on chronological age.

### Cell count measures: CD8T, CD4T, NK, Bcell, Mono, Gran

These are estimated proportions of CD8 T cells, CD4T cells, natural killer cells, B cells, monocytes and granulocytes. Toward this end, we used the method and R code described in Houseman et al (2012). Specifically, I used the R command "projectCellType" in the minfi R package (Aryee et al 2014). If you use these cell types in your work, make sure to cite Houseman et al 2014.

- *Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13:86. https://doi.org/10.1186/1471-2105-13-86*

### PlasmaBlast, CD8pCD28nCD45RAn, CD8.naive, CD4.naive

These are estimated abundance measures of plasma blasts, CD8+CD28-CD45RA- T cells, naive CD8 T cells, and naive CD4 T cells. Since a novel approach was used to arrive at these estimates, please cite the article (Horvath and Levine 2015) if you use these measures.

- *Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis. 2015; 212:1563–73. https://doi.org/10.1093/infdis/jiv277*

Interpretation: The resulting estimates should *not* be interpreted as counts or percentages but rather as ordinal abundance measures. Don't turn them into proportions (by dividing the measures by the sum). Negative values simply indicate very low values. Personally, I would not set a negative value to zero but would not object if you do that.

Biology:

a) CD8+CD28-CD45RA- T cells have characteristics of both memory and effector T cells. These cells increase with chronological age.

b) Naive CD8 T cells decrease with age.

c) Here naive CD8 and CD4 T cells are defined as CD45RA+CCR7+ cells.

d) Plasma cells, also called plasma B cells, and effector B cells, are white blood cells that secrete large volumes of antibodies. From Wikipedia: Upon stimulation by a T cell, which usually occurs in germinal centers of secondary lymphoid organs like the spleen and lymph nodes, the activated B cell begins to differentiate into more specialized cells. Germinal center B cells may differentiate into memory B cells or plasma cells. Most of these B cells will become plasmablasts, and eventually plasma cells, and begin producing large volumes of antibodies.

Statistical method for estimating these cell abundance measures: A penalized regression model (elastic net) was used regress cell count measures on DNA methylation levels. Estimated values are predicted values based on this penalized regression model.

propNeuron= proportion of neurons in nervous tissue calculate using the CETS algorithm

- Guintivano J1, Aryee MJ, Kaminsky ZA 2013 A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013 Mar;8(3):290-302. PMID: 23426267 PMCID: PMC3669121 DOI: 10.4161/epi.23924

DNAmTL= DNA methylation based estimator of leukocyte telomere lengt

- Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, Kooperberg C, Ferrucci L, Hou L, Baccarelli AA, Li Y, Harris SE, Corley J, Taylor A, Deary IJ, Stewart JD, Whitsel EA, Assimes TL, Chen W, Li S, Mangino M, Bell JT, Wilson JG, Aviv A, Marioni RE, Raj K Horvath S (2019) DNA methylation-based estimator of telomere length. Aging (Albany NY). 2019 Aug 18;11(16):5895-5923. doi: 10.18632/aging.102173. PMID: 31422385 PMCID: PMC6738410

## Additional details on the output file

Additional comments from the online software

Note that the output file contains a host of useful information e.g.

- SampleID=sample identifier
- DNAmAge=DNA methylation age=predicted age

- Comment=A comment is only added if a sample looks suspicious.
- noMissingPerSample=number of missing beta values per sample
- meanMethBySample, minMethBySample=the mean and min beta value before normalization
- corSampleVSgoldstandard=correlation between the sample and the gold standard (defined by averaging the beta values across the samples from the largest blood data set). A low value spells trouble and a comment will be added.
- meanAbsDifferenceSampleVSgoldstandard=mean absolute difference between the sample and the gold standard. A large value spells trouble and a comment will be added.
- predictedGender=predicted gender based on the mean across the X chromosomal markers. The sample is problematic if the predicted gender does not match the known gender.
- meanXchromosome= mean beta value across the X chromosomal markers. This variable is used for predicting gender. Female samples should have a higher value than male samples if X chromosomal inactivation is applicable.
- predictedTissue=the predicted DNA source (i.e. it does not have to be a tissue)
- ProbabilityFrom.Blood.PBMC=probability that the DNA derives from peripheral blood mononuclear cells.
- ProbabilityFrom.Brain.Cerebellar=probability that it comes from cerebellar brain samples
- ProbabilityFrom.Brain.FCTX=probability that it comes from frontal cortex
- ETC
- AgeAccelerationDiff=Age acceleration measure defined simply as difference, i.e. DNAmAge minus Age
- AgeAccelerationResidual=Age acceleration measure defined as residual from regressing DNAm age on chronological age. In R language: residuals(lm(DNAmAge-Age))

## Log file

The second email attachment (ending in log.txt) is a log file that briefly describes the data and provides some feedback, e.g. warnings or error messages.

### Cell count measures for multivariate regression models

Since cellular heterogeneity in blood can greatly affect DNAm studies, it is often a good idea to adjust for cell abundance measures.

The following imputed blood cell counts were analyzed: B cell, naïve CD4+ T, CD4+ T, naïve CD8+ T, CD8+ T, exhausted cytotoxic CD8+ T cells (defined as CD8 positive CD28 negative CD45R negative), plasma blasts, natural killer cells, monocytes, and granulocytes. The abundance of naive T cells, exhausted T cells, and plasma blasts were based on the Horvath method (Horvath and Levine 2015). The remaining cell types were imputed using the Houseman method. Toward this end, we use two types of blood cell counts

1. CD8.naive (Horvath method)
2. CD8pCD28nCD45RAn (Horvath method)
3. PlasmaBlast (Horvath method)
4. CD4T (Houseman)
5. NK (Houseman)
6. Mono (Houseman)
7. Gran (Houseman)

Since many of the cells are highly correlated with each other, I dropped the B cell and CD8T cell estimates from the Houseman method. When studying various diseases, it is probably a good idea to replace "Bcell" by "PlasmaBlast" (related to B cells) since the latter is often more disease relevant. Further, I usually replace "CD8T" by the two measures "CD8.naive"

"CD8pCD28nCD45RAn" since the latter are probably more disease relevant. I rarely use CD4.naive since CD8.naive is often more relevant.

To assess whether DNAmAge relates to a disease outcome, I use the following covariate list

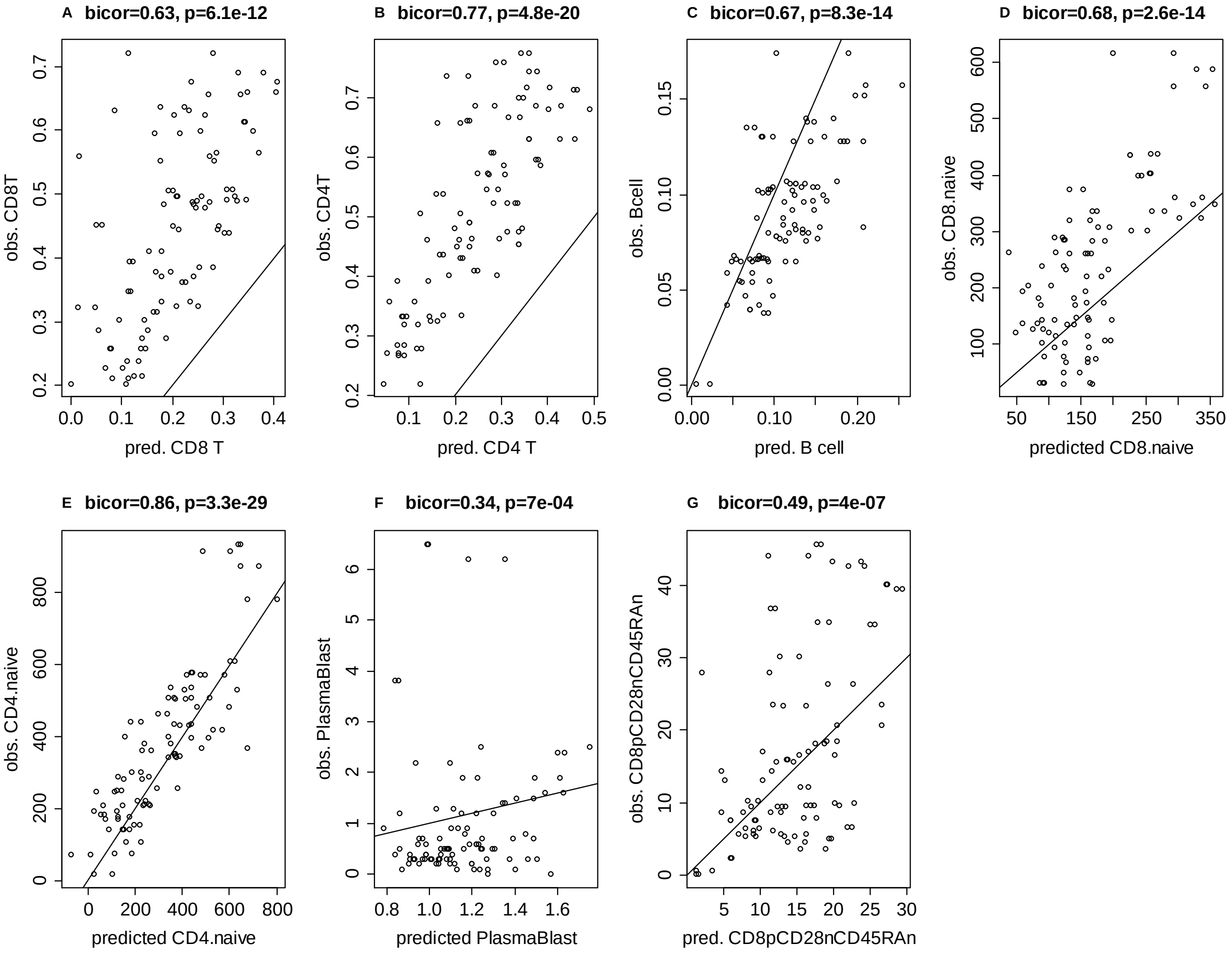
DNAmAge+Age+CD8.naive + CD8pCD28nCD45RAn + PlasmaBlast+CD4T+NK+Mono+Gran.

Obviously, you would also adjust for standard variables such as gender, race, body mass index, prior history of disease e.g. cancer, type II diabetes status, etc.

## How accurate are the cell count estimates produced by the epigenetic clock software?

The estimates are fairly accurate (median correlation=0.67, range=[.34,.86], Figure).

Details. I used an independent data set (which was not used to construct the cell count predictors). Using Illumina Inf450 data from 96 PBMC samples, we find fairly good agreement between the predicted cell abundance measure (x-axis) and the corresponding observed value (based on flow cytometry measures). Statistical detail: the reported correlation coefficients were calculated with the biweight midcorrelation, which is implemented in the WGCNA R function bicor. Panel E shows that the highest correlation (r=0.86) can be observed for naive CD4 T cells. The worst correlation can be observed for plasma blasts (r=0.34, panel F). The Houseman estimates for CD8T, CD4T, and B cells are presented in panels A-C. Panels D-G describe results for the above mentioned Horvath method.



## Why does the web based calculator not return any results for my data set?

Answer: A small data set (say fewer than 100 samples) should lead to a response within an hour or so. Try a small subset of your data to see whether you get a response. If not, your data lead to an error. Here are some common remedies.

a) If you uploaded a sample annotation file, make sure that its numbers of rows correspond to the number of samples, i.e. the numbers of columns of dat0 minus 1.

b) Make sure that your DNA methylation data file contains all the necessary probes. While it is OK to have missing DNA methylation levels, it is not OK to have missing probe IDs. Unless you use all probes on the 450K array or the 27K array, please make sure that your file includes ALL CpGs listed in datMiniAnnotation3.csv

Probes that were not measured in your data set should lead to a row filled with NAs. But the probe name needs to be listed. The advanced analysis option for blood requires that your data were measured on the Illumina450K platform but it only uses the probes in datMiniAnnotation.

c) Line feeds: I have noticed that the session breaks down when users upload the wrong line breaks. It should be CR+LF (carriage return and line feed) and not just LF or CR. A simple remedy is to open the csv file in Excel and save it as a .csv file for Windows.

d) Make sure that you upload numeric data (missing values should be coded as NA and not as null or NULL). Sometimes a user uploads a file that also contains various annotations (e.g. chromosome number, gene name). Carefully look at dat0 before you upload it. The first column should contain CpG identifiers. The remaining columns should only contain numeric values. If a column (sample) only contains missing values, remove it from dat0 and datSample. If need be, run the following R code before you upload the data.

```
for (i in 2:dim(dat0)[[2]] ) { dat0[,i]=as.numeric(as.character(dat0[,i])) }
```

## Frequently asked questions

### Q: Does the order of the samples in the sample annotation file have to match that of the methylation file?

A: Yes, absolutely. If DNAm age is not correlated with chronological age then there is a good chance that the user or the lab accidentally permuted the sample order. I could tell you several anecdotes about how the epigenetic clock software allowed us to find plating errors or labeling errors.

### Q: Are additional columns allowed in the sample annotation file?

Yes, as many as you can handle. Thousands.

### Q: Does the order of the columns matter in the sample annotation file? It seems like you will require the first column to be "SampleID", second column "Age".

A: No the order does not matter. The first column does *not* have to be called SampleID. However, it is very important that the file contains columns called "Age", "Female", and "Tissue". The capitalization has to be as specified. Don't use variable names such as age, AGE, female, tissue, TISSUE.

## References

- Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115 PMID: 24138928
- Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013 Jan 24;49(2):359-67.

- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012, 13:86  doi:10.1186/1471-2105-13-86
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014 May 15;30(10):1363-9. doi: 10.1093/bioinformatics/btu049.