

TUTORIAL for the Online Age Calculator:

Estimate DNA methylation age

Steve Horvath (shorvath at mednet.ucla.edu)

This tutorial illustrates how to calculate DNA methylation age using the online calculator.

Mandatory input: A (compressed) file with beta values, e.g. measured on the Illumina 27k or 450k platform. 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.

Output:

- DNAMAge=predicted age (referred to as DMAM age)
- corSampleVSGoldstandard quality statistic for detecting outlying samples (e.g. corSampleVSGoldstandard<0.8 should probably be excluded)

Optional, additional input: I recommend that you also input a sample annotation file that specifies age, tissue, etc. In this case use the following variable names "**Age**" (note it starts 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 for Blood

If you applied the Illumina 450K platform to blood then you can get a host of additional output by selecting the AdvancedAnalysisBlood option. In this case, the software will output

- additional measures of biological age in blood
- estimates of blood cell counts
- different measures of age acceleration.

Citation of this software

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

Contents

How to upload the data?	4
Upload form	5
Strategies for uploading very large data sets	5
Normalization, imputation.....	6
Uploading the sample annotation file	7
After you push the submit button	8
Output file	9
Log file	10
Advanced Analysis in Blood	10
1) BioAge1HO, BioAge2HO, BioAge3HO, BioAge4HO.....	10
2) BioAge1HA, BioAge2HA, BioAge3HA, BioAge4HA.....	10
3) BioAge2HOStatic, BioAge3HOStatic, BioAge4HOStatic	10
4) BioAge2HASstatic, BioAge3HASstatic, BioAge4HASstatic	11
5) BioAge1HOAdjAge, BioAge2HOAdjAge, BioAge3HOAdjAge, BioAge4HOAdjAge.....	11
6) BioAge1HAAdjAge, BioAge2HAAdjAge, BioAge3HAAdjAge, BioAge4HAAdjAge	11
7) BioAge2HOStaticAdjAge, BioAge3HOStaticAdjAge, BioAge4HOStaticAdjAge	11
8) BioAge2HASstaticAdjAge, BioAge3HASstaticAdjAge, BioAge4HASstaticAdjAge	11
9) PlasmaBlastAdjAge, CD8pCD28nCD45RAnAdjAge, CD8.naiveAdjAge, CD4.naiveAdjAge	11
10) Cell count measures: CD8T, CD4T, NK, Bcell, Mono, Gran	11
11) PlasmaBlast, CD8pCD28nCD45RAn, CD8.naive, CD4.naive	12
12) Cell count measures for multivariate regression models.....	12
12) AAHOAdjCellCounts and AAHAAdjCellCounts	13
How accurate are the cell count estimates produced by the epigenetic clock software?	13
Why does the web based calculator not return any results for my data set?	14
Frequently asked questions	15
Q: Does the order of the samples in the sample annotation file have to match that of the methylation file?	15
Q: Are additional columns allowed in the sample annotation file?	15
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"	15

Q: In the "Advanced Analysis in Blood" option, the 4 weighted averages are a bit of a mystery as currently described. Can you elaborate on how the weighted averages were calculated?..... 15

Q: In the advanced analysis option, it appears that only 2 age acceleration measure account for cell types (e.g. "AAHOAdjCellcounts" and "AAHAAdjCellcounts"). Which epigenetic age measure is being used? 15

References 16

Instructions

Go to the webpage: <http://labs.genetics.ucla.edu/horvath/dnamage/>

To run this tutorial, download the following example data set from the webpage

MethylationDataExample55.csv

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.

ProbeID	GSM946048	GSM946049	GSM946052	GSM946054	GSM946055	GSM946056	GSM946059	GSM946062	GSM946064	GSM946065	GSM946066	GSM946067
cg00000292	0.705861431	0.729790365	0.704587006	0.750851623	0.715302226	0.63440712	0.681816613	0.635320371	0.727532164	0.777193913	0.601302367	0.75166
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
cg00003994	0.037032466	0.014692376	0.017115722	0.028960542	0.016254482	0.024279664	0.012705737	0.020628148	0.015136485	0.010465486	0.028988007	0.01445
cg00005847	0.133246824	0.12048361	0.120860048	0.106939397	0.109781934	0.12935425	0.102437983	0.123934306	0.104037878	0.108495207	0.12248214	0.12528
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
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
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
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
cg00009407	0.010457567	0.005184467	0.00410315	0.006713128	0.007575285	0.005175612	0.005425126	0.006243924	0.006421381	0.006799044	0.00712126	0.00765
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
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
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
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
cg00012792	0.025206845	0.025953047	0.031571295	0.025945964	0.027536842	0.023158617	0.025262766	0.025490522	0.02561854	0.023987126	0.029929646	0.03258
cg00013618	0.945071544	0.934224335	0.935316359	0.940588786	0.938782958	0.933117967	0.927241412	0.928640055	0.932432338	0.926540548	0.937869607	0.92604
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
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
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
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
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
cg00020533	0.963626981	0.97589055	0.974573375	0.970928117	0.974559639	0.972896393	0.965543018	0.967269887	0.957033077	0.965032624	0.965421601	0.94592
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
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
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
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
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
cg00025138	0.01110685	0.008542402	0.006302831	0.009048799	0.007566799	0.009319309	0.007477717	0.007287072	0.011238628	0.010465121	0.007966301	0.00837
cg00025991	0.712835775	0.700209783	0.648668303	0.60936276	0.63247454	0.683824411	0.672282499	0.673337413	0.623947753	0.617621511	0.680760397	0.60337

In this tutorial, I analyze data set 55:

- 16 men: autistic subjects and controls

- brain occipital cortex samples
- Illumina 27K platform
- GEO accession GSE38608
- Citation for the data set:

Ginsberg MR, Rubin RA, Falcone T, Ting AH et al. Brain transcriptional and epigenetic associations with autism. PLoS One 2012;7(9):e44736. PMID: [22984548](https://pubmed.ncbi.nlm.nih.gov/22984548/)

Some comments for the experts:

These DNA methylation data were downloaded from the Gene Expression Omnibus data base (GEO accession GSE38608). GEO allows users to post both normalized data and raw data. The authors posted M values as normalized values. However, my age predictor makes use of beta values since I did not find any evidence that M values are superior to beta values when it comes to age prediction.

Message: the beta values used in this tutorial do not match the normalized (M value) data from GEO. But it is straightforward to turn M values into beta values...

How to upload the data?

Note that the following webpage <http://labs.genetics.ucla.edu/horvath/dnamage/>

Steve Horvath (shorvath at mednet.ucla.edu)

DNA methylation age and the epigenetic clock

Abstract

Using 82 Illumina DNA methylation array data sets (n=7844) involving 51 healthy tissues and cell types, I developed a multi-tissue predictor of age which allows one to estimate the DNA methylation (DNAm) age of most tissues and cell types. DNAm age has the following properties: a) it is close to zero for embryonic and induced pluripotent stem (iPS) cells, b) it correlates with cell passage number, c) it gives rise to a highly heritable measure of age acceleration, and d) it is applicable to chimpanzee tissues.

Frequently asked questions (FAQs): [faq.htm](#)

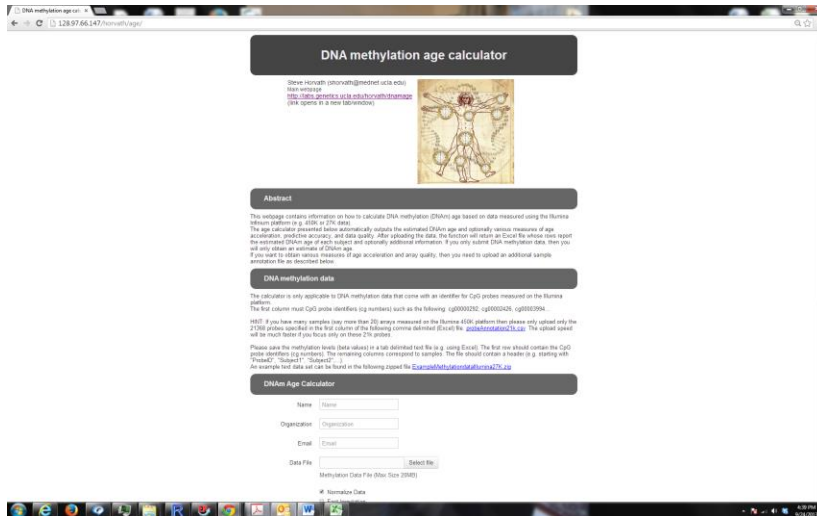
DNAm age calculator

DNAm age calculator webpage (below) contains information on how to calculate DNA methylation (DNAm) age based on data measured using the Illumina Infinium platform (e.g. 450K or 27K data). The age calculator automatically outputs the estimated DNAm age and optionally various measures of age acceleration, predictive accuracy, and data quality. After uploading the data, the function will return an Excel file whose rows report the estimated DNAm age of each subject and optionally additional information. If you only submit DNA methylation data, then you will only obtain an estimate of DNAm age. If you want to obtain various measures of age acceleration and array quality, then you need to upload an additional sample annotation file as described below.

[Access Online Age Calculator](#)

contains a hyperlink called "[Access Online Age Calculator](#)".

After you click it you will arrive at the following webpage



Upload form

In the online form, enter your

1. Name:
2. Organization:
3. **Email address.** 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 450K data into a "reduced" file that only contains probes that can be found in the file `datMiniAnnotation.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.

```

library(sqldf)

#change the setwd filepath to that of the folder with your data. Note the forward slash
setwd("C:/Users/SHorvath/Documents/DNAMAge/Example55")

#replace "MethylationData.csv" with the name of your methylation data file
dat0=read.csv.sql("MethylationData.csv") ;

datMiniAnnotation=read.csv("datMiniAnnotation.csv")

match1=match(datMiniAnnotation[,1], dat0[,1] )

dat0Reduced=dat0[match1,]

dat0Reduced[,1]=as.character(dat0Reduced[,1])

dat0Reduced[is.na(match1),1]= as.character(datMiniAnnotation[is.na(match1),1])

datout=data.frame(dat0Reduced)

# make sure you output numeric variables...

for (i in 2:dim(datout)[[2]] ){datout[,i]=
as.numeric(as.character(gsub(x=datout[,i],pattern="",replacement=""))) }

#replace "MethylationData" with a filename of your choice

write.table(datout,"MethylationData.csv", row.names=F, sep="," )

```

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

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 file

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.

Advanced Analysis in Blood

If you measured Illumina 450K data in blood then I recommend that you select the advanced analysis option in blood. Side note: If you have more than say 100 samples then I strongly recommend to use data compression strategies 2 and 1 described in [Strategies for uploading very large data sets](#).

The advanced analysis option leads to a host of additional output: various measures of biological age, age acceleration and blood cell counts.

1) BioAge1HO, BioAge2HO, BioAge3HO, BioAge4HO

Explanation: All of these measures of biological age generalize the DNAmAge described in Horvath 2013. BioAge1HO is a "calibrated" version of DNAmAge (which means that a linear transformation was applied so that it is aligned with chronological age). BioAge2HO, BioAge3HO, BioAge4HO are defined as weighted average based on two, three, and four epigenetic input variables, respectively. The weights are "dynamically" calculated by correlating the input variables to chronological age. Measures 2-4 can only be calculated if chronological age specified in the variable "Age" is available and has a non-zero variance. If age is not available or all samples have the same age (zero variance) simply use [3\) BioAge2HOStatic, BioAge3HOStatic, BioAge4HOStatic](#) .

2) BioAge1HA, BioAge2HA, BioAge3HA, BioAge4HA

Explanation: All of these measures extend the predicted age measures based on the 71 CpGs in Hannum 2013. BioAge1HA is a "calibrated" version of the predicted age from Hannum. The calibration was defined as a linear transformation that ensures that the predicted age is aligned with chronological age. Again, BioAge2HA, BioAge3HA, BioAge4HA are generalized measures of biological age based on two, three, and four epigenetic input variables, respectively. They can only be calculated if chronological age specified in the variable "Age" is available. If age is not available or all samples have the same age (zero variance) simply use [4\) BioAge2HASstatic, BioAge3HASstatic, BioAge4HASstatic](#)

3) BioAge2HOStatic, BioAge3HOStatic, BioAge4HOStatic

These measures are analogous to those described in [1\) BioAge1HO, BioAge2HO, BioAge3HO, BioAge4HO](#) but the weights are static (meaning constant). In particular, these measures can be calculated even if the column "Age" is filled with missing values or all subjects have the same chronological age.

4) BioAge2HStatic, BioAge3HStatic, BioAge4HStatic

These measures are analogous to those described in [2\) BioAge1HA, BioAge2HA, BioAge3HA, BioAge4HA](#) but the weights are static (meaning constant). In particular, these measures can be calculated even if the column "Age" is filled with missing values or all subjects have the same chronological age.

5) BioAge1H0AdjAge, BioAge2H0AdjAge, BioAge3H0AdjAge, BioAge4H0AdjAge

These are measures of age acceleration based on adjusting measures [1\) BioAge1HO, BioAge2HO, BioAge3HO, BioAge4HO](#) for chronological age using a linear regression model.

In other words, a positive (negative) value indicates that the biological age of the sample is higher (lower) than expected based on chronological age. Specifically, these measures were defined as residuals (observed minus predicted) resulting from a simple linear regression model that regressed biological age on chronological age. By definition, these measures are not correlated with chronological age, which is an attractive property. A disadvantage of these measures is that they can only be defined if "Age" has a non-zero variance and is available for at least 4 subjects.

6) BioAge1HAAdjAge, BioAge2HAAdjAge, BioAge3HAAdjAge, BioAge4HAAdjAge

These are measures of age acceleration based on adjusting measures [2\) BioAge1HA, BioAge2HA, BioAge3HA, BioAge4HA](#) for chronological age using a linear regression model.

7) BioAge2H0StaticAdjAge, BioAge3H0StaticAdjAge, BioAge4H0StaticAdjAge

These are measures of age acceleration based on adjusting measures [3\) BioAge2H0Static, BioAge3H0Static, BioAge4H0Static](#) for chronological age using a linear regression model.

8) BioAge2HStaticAdjAge, BioAge3HStaticAdjAge, BioAge4HStaticAdjAge

These are measures of age acceleration based on adjusting measures [4\) BioAge2HStatic, BioAge3HStatic, BioAge4HStatic](#) for chronological age using a linear regression model.

9) PlasmaBlastAdjAge, CD8pCD28nCD45RAnAdjAge, CD8.naiveAdjAge, CD4.naiveAdjAge

These are age adjusted versions of [11\) 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.

10) 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, I 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.

11) PlasmaBlast, CD8pCD28nCD45RA, 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 epigenetic clock software (Horvath 2013) if you use these measures. 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.

12) Cell count measures for multivariate regression models

Since cell heterogeneity can greatly affect DNAm studies, it is often a good idea to adjust for cell abundance measures (Houseman 2012). If you have whole blood methylation data, then I suggest you adjust for the following cell counts in your analysis:

1. CD8.naive (Horvath method)
2. CD8pCD28nCD45RA (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.

12) AAHOAdjCellCounts and AAHAAdjCellCounts

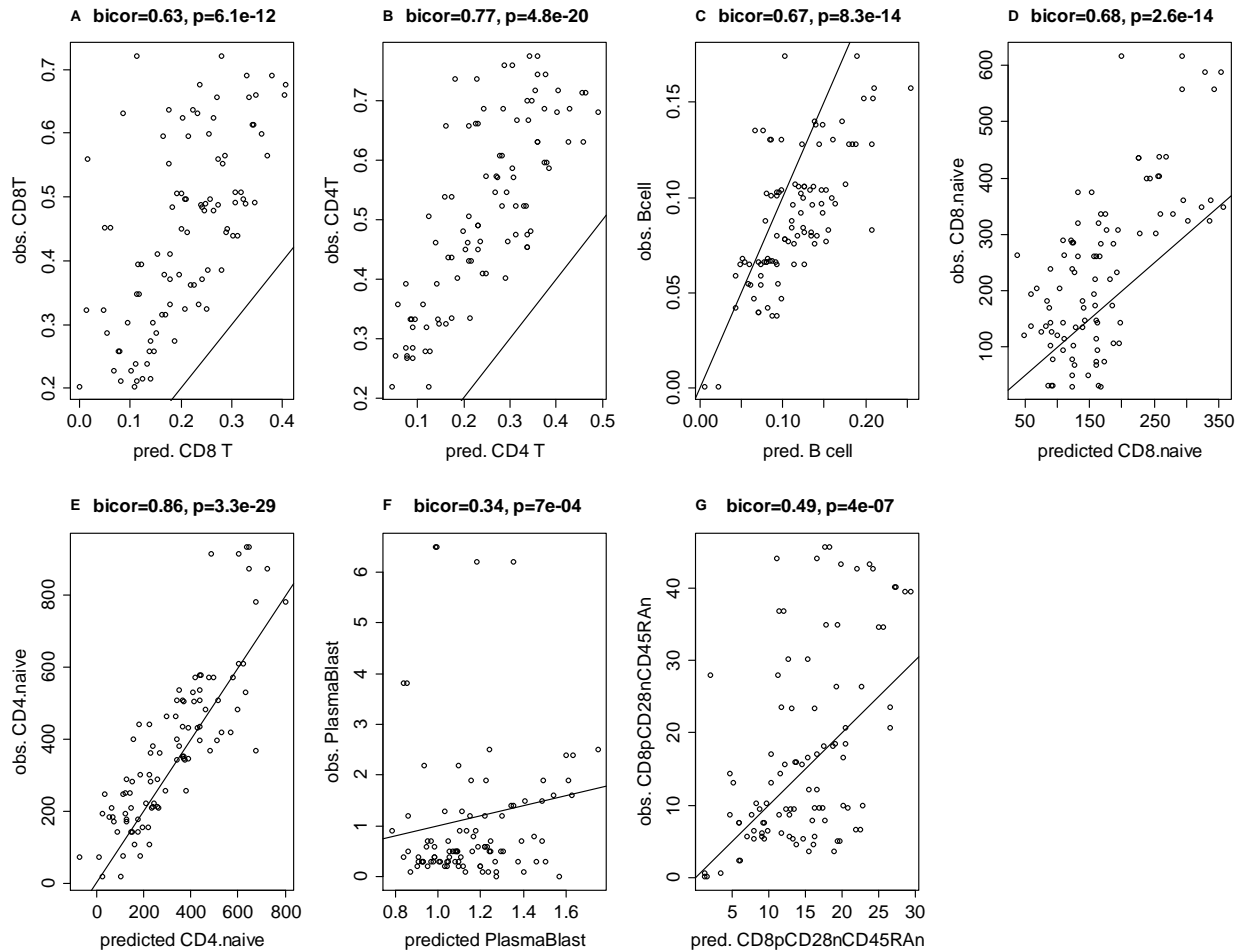
These are measures of age acceleration that adjust for cell counts. Specifically, these are residuals resulting from multivariate regression models that regress an estimate of DNAm age on age+CD8.naive + CD8pCD28nCD45RAn + PlasmaBlast+CD4T+NK+Mono+Gran (as described in the previous section).

AAHOAdjCellCounts and AAHAAdjCellCounts correspond to age acceleration measures based on Horvath (2013) and Hannum (2013), respectively. It turns out that each of the following four measures [1\) BioAge1HO](#), [BioAge2HO](#), [BioAge3HO](#), [BioAge4HO](#) leads to the same measure AAHOAdjCellCounts. Similarly, each of the following four measures [2\) BioAge1HA](#), [BioAge2HA](#), [BioAge3HA](#), [BioAge4HA](#) leads to the same measure AAHAAdjCellCounts.

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 [datMiniAnnotation.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.

Q: In the "Advanced Analysis in Blood" option, the 4 weighted averages are a bit of a mystery as currently described. Can you elaborate on how the weighted averages were calculated?

These measures have not yet been published. Please email Steve Horvath to request a relevant manuscript.

Q: In the advanced analysis option, it appears that only 2 age acceleration measure account for cell types (e.g. "AAHOAdjCellcounts" and "AAHAAdjCellcounts"). Which epigenetic age measure is being used?

A: It turns out that multiple measures of biological age lead to the same adjusted measure, which is why it is sufficient to calculate only two age acceleration measures that account for cell counts. Specifically, each of the following four measures 1) BioAge1HO, BioAge2HO, BioAge3HO, BioAge4HO leads to the same measure AAHOAdjCellCounts. Similarly, each of the following four measures 2) BioAge1HA, BioAge2HA, BioAge3HA, BioAge4HA leads to the same measure AAHAAdjCellCounts.

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, Sada 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.