International Journal of Genetic Engineering and Biotechnology.
ISSN 0974 3073 Volume 6, Number 1 (2017), pp. 1-12

http://www.irphouse.com

Age estimation using neural networks and DNA methylation levels

Gerardo Alfonso
University Autonoma Barcelona
Genetics Department

Abstract

There are several techniques available to determine the biological age of a patient by analyzing the DNA methylation levels of some of their cells. In this article the forecasting accuracy of neural networks is compared to the k-nearest neighbors ("KNN") technique. The accuracy of the forecast is related to the sample size. For smaller datasets the KNN provide some moderately accurate results, with an average error of approximately 10 years. When the sample size increase the KNN does not appear to work properly (for a dataset of 720 samples) and neural networks start to provide better results. While the amount of samples in each dataset varied the number of CpGs per case was constant at approximately 27,000. Several simulations were performed randomly reducing the number of CpGs in the samples. It was found that typically the best results were found not when using all the CpGs (27,000) but when using a relatively randomly selected subset of approximately 300 to 400 CpGs. While the sample size is too small to be conclusive the results seem to indicate that for an age forecasting point of view a significant fraction of the CpGs methylation data might add mostly noise. It is clearly required further work to determine is this intuition is actually correct.

Keywords: aging, DNA methylation

INTRODUCTION

DNA methylation is a normal process that is impacted by environmental factors [1] and life style. Indications of the central role that methylation has in many biological process is known since 1979, after the highly influential paper by McGhee [2]. Since then there has been an increasing amount of literature involving methylation and many specific processes from diseases [3], [4], [5], [6] to aging [7]. From a biochemical point of view methylation occurs when a methyl group links to a base (either C or G). The level of methylation changes from tissue to tissue and it is different if the individual has some disease, such as cancer. Age is also a factor impacting methylation. There are indications that newborn methylation levels are impacted by maternal smoking during pregnancy [8] and even might have an impact on memory [9]. Currently it is relativity straightforward obtaining methylation data from many different cells, such as sperm [10] or colon cells [11] with the majority of the sample publically available being of whole blood. One widely accepted technique to determine DNA methylation levels is bisulphite modification [12]. Thank you to this technique and similar approaches the accuracy of DNA methylation measurements has increased substantially over the last decade. While there has been a large amount of research regarding methylation there continuous to be many questions remaining such as the exact role the methylation has in the aging process or if methylation changes can be induced to prevent certain illnesses.

METHYLATION AND AGING

Methylation has been mentioned in a multitude of research reports as an aspect influencing the aging process in humans [13]. Changes in methylation levels related to aging have been measured not only in humans but also in some other species such as mice [14], salmon [15], [16] or great apes [17]. There seems to be a consensus in the literature with the existence of some type of relationship between DNA methylation levels and aging but less of a consensus of how the aging process actually occurs or if changes in DNA methylation can actually increase life spans [15]. Abnormal methylation levels do appear to be related with premature aging and some illnesses. It should be noted that currently it is possible to induce changes in DNA methylation [18] and that this is an active area of research. Methylation alteration has been mentioned as an easier way to modify DNA than through mutations [6].

There currently exist accurate multi tissue clocks, such as [19], that can predict biological age of a person using methylation levels from several different types of

tissues with an error of only a few years. All these indications points towards some type or relationship between methylation levels and aging and warrant doing further research on what statistical applications to use. In this article neural networks and the k- nearest neighbor approach were followed to link those DNA methylation levels with the patient age.

NEURAL NETWORKS

Neural networks are a statistical application that has proven valuable for signal fitting. It is biologically inspired and similar to many other machine learning application does not require theoretical knowledge of the relationship between the input and the output. The first theoretical steps in the neural network space track back to the late 50^{th} early 60^{th} but these techniques only became popular several decades later with the development of computers. One of the most successful applications of neural networks was in the field of supervised learning. For supervised learning applications a neural network composed of a number of neurons is trained to replicate an actual output as closely as possible by adjusting the relative importance of the value of those neurons. Then the network is typically tested with new data to try to identify its generalization power. There is a huge amount of different neural networks. Some of the main differences are the network structure, the type of neurons used and the training algorithm. Neural networks have been applied in some areas of medical research, such as forecasting of growth of staphylococcus in milk [20] or medical diagnosis [21]. These techniques are typically used when the underlying relationship between the exogenous and endogenous variables is not known or when such relationship is too complicated to model explicitly.

K-Nearest Neighbors (KNN)

The k nearest neighbor ("KNN") technique is a frequently used [22] statistical classification tool with application in many different areas such as image interpolation [23], heart disease evaluation [24] or engine diagnosis [25]. KNN is a non-parametric test. The basic idea of the algorithm is to find the k nearest neighbors from a set of inputs $(x_1, \dots, x_n)$ that has the smallest distance from a given set of outputs. There is no theoretical upper limit in the number of neighbors (k) that can be used in the algorithm with the optimal amount determined by the specific characteristics of the problem [26].When k=1 the algorithm is called nearest neighbor. KNN is an intuitive

approach useful with inputs of high dimensionality as the ones related to methylation analysis. The approach has been proven in many applications that needed filtering the input information for irrelevant data [27] and can handle prediction between the input data x and the output data y without needing to now the function $f(x) = y$ that relates both variables. There are also several measures of distance that can be used in the algorithm with Euclidean distance among the most frequently used. The distance measures used in this article can be found in (table 1). KNN is a supervised learning technique. In a first instance the training input data (x) are classified into categories according to the training value. Then a new set of inputs are used for testing purposes. The nearest neighbor for the new testing data is found and then categorized accordingly. The error is the difference between this categorization and the actual category that the data belongs to.

KNN while able to provide accurate forecasts for many applications [28], [29], [30] have their own drawbacks. Some of the most frequently mentioned in the literature are the substantial computational requirements [31], with a tremendous amount of distances needed to be calculated when canalizing a large data set like those typically related to DNA methylation and the related memory requirement [31]. These disadvantages, while clearly important, are less relevant that in other applications that require real time functionality.

Table 1. Distance measures

Measure	Definition
Euclidean	$d = \sqrt{(x - p)}$
City block	$d = \lvert x - p \rvert$
Correlation	$d = 1 - \text{correlation}(x, p)$
Cosine	$d = 1 - \text{angle}(x, p)$

Compared to other topics there is relatively little research done on applications of KNN in DNA methylation applications. [32] is an interesting article using the KNN in methylation data for gene expression. The authors find in this article that the KNN approach generated better classification results for breast cancer determination than other commonly used algorithms. Another interesting article in this regard is [33] that

used support vector regression models ad KNNs for interpolating some missing data points. KNN tend to be used in the literature for such type of filling missing data than for the actual estimation.

METHODOLOGY

Three databases of different sizes were analyzed in this article. The following databases were used in this article: 1) (GSE56606) containing methylation information of CD14+ monocytes for 100 patients with diabetes as well as control subjects from an article by Rakyan, [34], 2) (GSE34035) containing methylation data for saliva of 197 patients with different alcohol consumption from an article by Liu [35], 3) (GSE24884) methylation data of subcutaneous adipose tissue of 56 patients from an article by Arner [36], and 4) (GSE41037) database of 720 patients suffering from schizophrenia as well control subjects from an article by Horvath [37]. All the datasets are publically available.

In a first instance, a neural network (backpropagation) with 10 neurons was applied to all the three datasets. As expected the results were considerably more accurate for the large data set than for the smaller ones. No meaningfully prediction was obtained for the smaller datasets using neural networks.

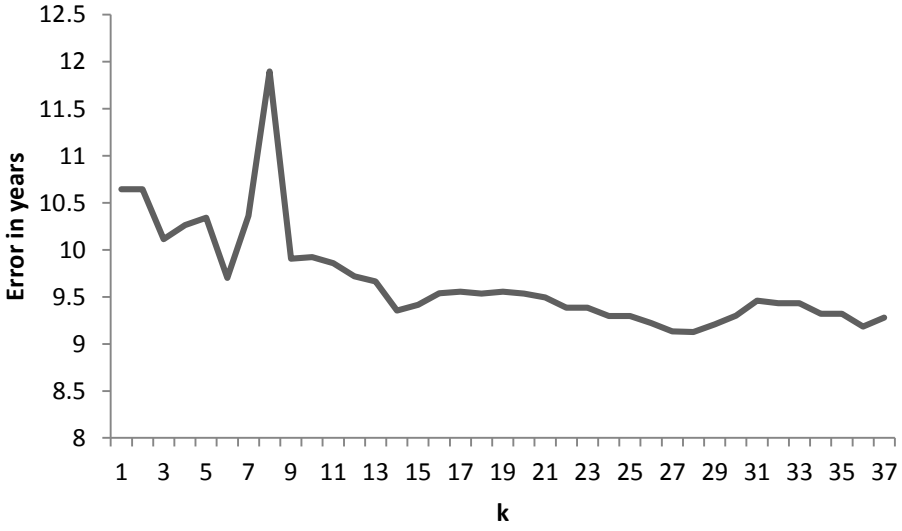

Figure 1: Mean error for out of sample values - KNN (GSE34035)

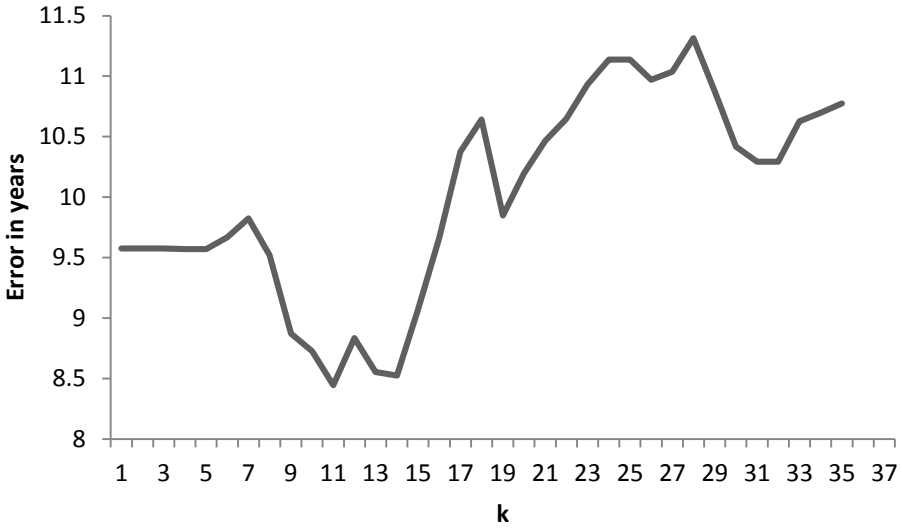

Figure 2: Mean error for out of sample values - KNN (GSE4996)

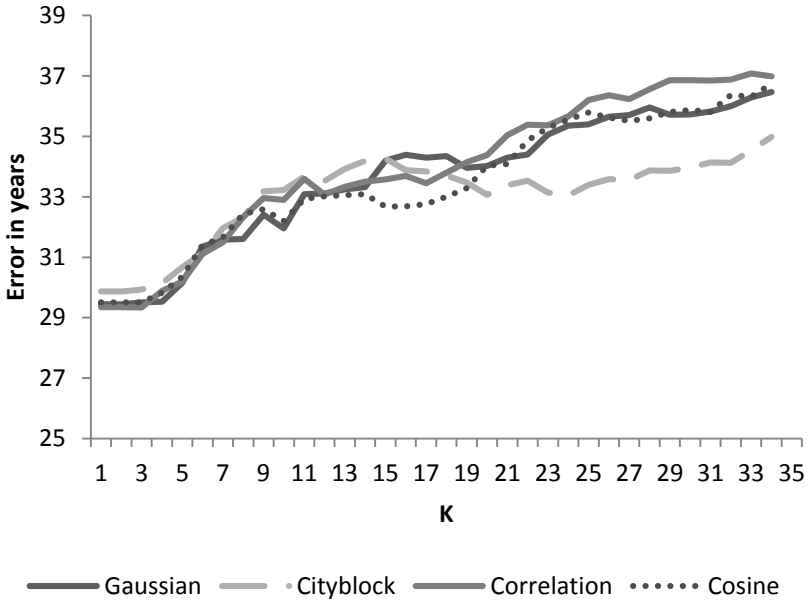

Figure 3: Mean error for out of sample values - KNN (GSE41037)

Then the KNN technique was used. The KNN approach was followed using a series of values of k, from 1 to 35 as, well as for several distance measures, such as Euclidean, Cityblock, Correlation and Cosine. The results, for some of the smaller datasets, can be seen in figures 1 and 2. The GSE41037 dataset was then sliced into smaller subsets to see if as the number of samples increased the accuracy of the KNN improved (figure 3) but this was not observed (perhaps due to the limited sample size). The regressions for two subsets of GSE41037 can be seen in figures 3 (75 cases) and 4 (100 cases) and some more details in table 2.

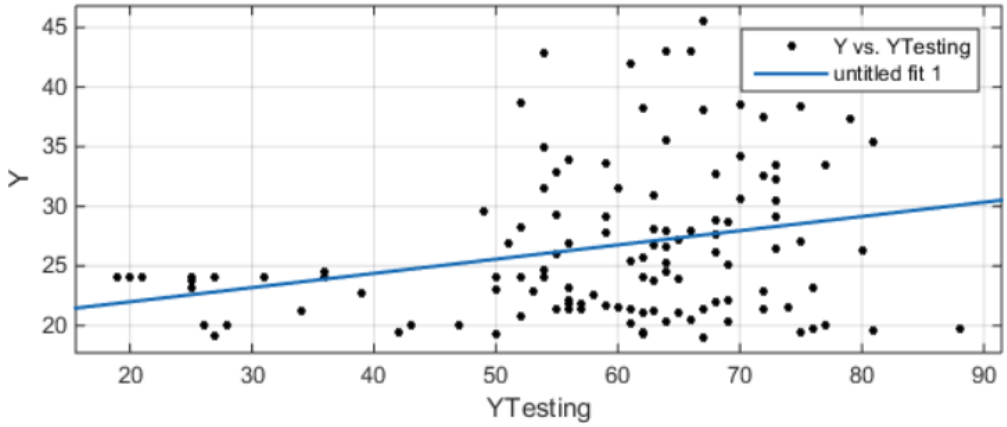

Figure 4: Linear regression with 75 cases (GSE41037 subset)

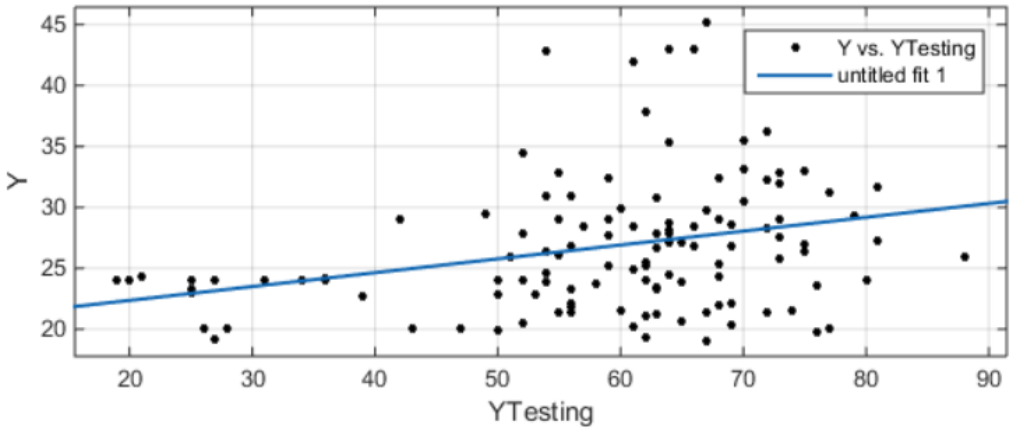

Figure 5: Linear regression with 100 cases (GSE41037 subset)

Table 2. Regression model f(x) = P1*x + P2 for various amount of samples (GSE41037)

# of samples	75	100	575
P1*	[0.0042,0.1969]	[0.0467,0.1806]	[0.0779,0.2388]
P2*	[14.89,24.33]	[16.02.24.17]	[19.11,28.90]
R square	0.07279	0.08735	0.1142

*95% confidence interval

The sensitivity of the results regarding the number of CpGs included in the analysis was also tested. The dataset GSE24884 was first analyzed with neural networks including all the CpGs in the dataset. 15 neural networks were performed in order to determine a median value for R and a probability distribution. As expected, due to the small amount of sample this approach did not produced an accurate forecast, with the mean value coming at -0.120. Then 50% of randomly selected CpGs were deleted from the data set and the process repeated reducing the amount of CpGs by approximately 50% in each step. This was performed iteratively until only approximately 100 CpGs were left. The best forecast obtained in this way was when using approximately 300 CpGs randomly selected, with a mean value of 0.292. This process was repeated 20 times generating each times a different subset of CpGs to be deleted. For all the 20 subsets, except one, the best combination was when using approximately 300 CpGs.

RESULTS

The techniques showed in this article, both neural networks and KNN, need a certain minimum amount of data to function properly but the sensitivity to the actual number of sample appear to be rather different. If the data set is relatively small the results seem to show that KNN works moderately well, regardless of the distance metric used with mean errors of approximately 10 year, while for larger datasets the neural network approached seemed to work better for the analyzed cases.

The mean error found using the KNN approach is not smaller than the one found by some other researchers but given the rather small sample size it is a reasonable result. The KNN approach seemed to produce values that were moderately sensitive to k within the specified range with a maximum difference in the error of approximately four years. For some datasets, such as GSE3403, increasing the value of k seemed to

decrease the error (figure 1) but this was not a constant trend for all the datasets analyzed. In fact for some datasets, such as (GSE4996), the error seemed to increase after a certain value of k (figure 2). The average errors (over all the k values) were 9.67, 10.01 and 9.87 years for the datasets (GSE34035), (GSE49996) and (GSE 24884). The error for the larger dataset (GSE41037) was actually large than for the smaller ones, coming at 33.61, 33.03,33.93 and 33.49 year using the Euclidean, Cityblock, Correlation and Cosine distance metrics respectively.

For small data samples the neural network approach did not seem to produce accurate forecasts. Forecasting accuracy did increase as the number of samples increased with the R value for the larger database using just 10 neurons coming at a reasonable average of 0.63 for the GSE41037 dataset. The accuracy of the neural network forecast, while changing the amount of CpGs included in the simulations, were analyzed. For the dataset analyzed the best results were obtained not when using all the CpGs but when using a relatively small amount of approximately 300 CpGs selected randomly. For instance, for the small dataset (GSE 24884) the best results obtained was for a subset of 354 CpGs with the 95% confidence interval for R being [0.0514, 0.5330], which was the only entirely positive interval for the combinations analyzed. It is important to keep in mind that a poor result for small datasets was expected. For the large dataset (GSE41037) the best result, such as the previously mentioned average 0.63 was obtained also with a subset of approximately 300 CpGs. More research is needed to explore this issue but the results seem to support the idea that a large amount of the CpGs might add mostly noise for age calculation purposes. It is also interesting that the results seem to be relatively consistent even when taking several randomly selected sets of 300 CpGs.

REFERENCES

1) Lam Lucia, Emberly Eldon, Hunter Fraser et al, 2012. “Factors underlying variable DNA methylation in human community cohort”. Proceedings of the national academy of sciences of the United States of America. Vol 109.

2) McGhee J, Ginder GD, 1979. “Specific methylation sites in the vicinity of the chicken beta globin genes”. Nature.

3) Cerchietti Leandro, Melnick Ari, 2017, “DNA methylation-based biomarkers. Journal of clinical oncology”. Vol 36, No. 7.

4) Nordlund J, Bäcklin C, Wahlberg P, Busche S, 2013, “Genome wide signatures

of differential DNA methylation in pediatric acute lymphoblastic leukemia". Genome Biology. Vol 14.

5) Shi J, Marconett C, Duan J, Hyland P, 2014. "Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue". Nature Communications. Vol 5.

6) Simons Daniels, 2008, "Epigenetic influences and disease". Nature education, Vol 1.

7) Mansego Maria Luisa, Milagro Fermin, Zulet Maria Angeles, Moreno-Aliaga Mari et al, 2015. "Differential DNA methylation in relation to age and health risk obesity". International journal of molecular sciences. Vol 16.

8) Bonnie Joubert, Janine Felix, Yousefi Paul, Bakulski Kelly et al, 2016. "DNA methylation and maternal smoking in pregnancy: Genome-wide consortium meta-analysis". The American journal of human genetics. Vol 98.

9) Day Jeremy, Sweatt Davis, 2010, "DNA methylation and memory formation. Nature neuroscience". Vol 13.

10) Cassuto Nino, Montjean Debbie, Siffroi Jean-Pierre, Bouret Dominique, 2016, "Different levels of DNA methylation detected in human sperms after morphological selection using high magnification microscopy". Biomedical research international.

11) Fernandez A, Martin Subero JI, Balint B, 2012, "DNA methylation fingerprinting of 1628 human samples". Genome research. Vol 22.

12) Patterson Kate, Molloy Laura, Qu Wenjia, Clark Susan, 2011, "DNA methylation: bishulphite modification and analysis". Journal of visualized experiments. Vol 56.

13) Rowbotham David, Marshal Erin, Vucic Emily, 2014, "Epigenetic changes in aging and age-related disease". Journal of aging science. Vol 3.

14) Langie Sabine, Cameron Kerry, Ficz Gabriella, Oxley David, Tomaszewski Bartlomiej et al, 2017, "The aging brain: effects on DNA repair and DNA methylation on mice". Genes. Vol 8.

15) Jung Marc, 2015, "Aging and DNA methylation". BMC biology. Vol 13.

16) Berdishev M, Korotaev G K, Boiarskikh G, 1967, "Nucleotide composition of DNA and RNA from somatic tissues of humpback salmon and its changes

during spawning". Biokhimiia. Vol 38.

17) Hernando-Herraez Irene, Padro-Martinez Javire Padro. Garg Paras, Fernandez-Callejo Marcos et al, 2013, "Dynamics of DNA methylation in recent human and great ape evolutions". PLoS.

18) Egger C, 2004, "Epigenetics in human disease and prospects for epigenetic therapy. Nature". 429.

19) Horvath S, 2013, "DNA methylation age of human tissues and cell types. Genome biology". Vol 14.

20) Orawan C, Pandawee S, Bandit S, 2016, "Application of artificial neural networks on growth prediction of staphylococcus in milk". International food research journal. 23(1).415-418.

21) Qeethara Kadhim, 2011, "Artificial neural networks in medical diagnosis". IJCSI International journal of computer sciences issues. Vol 8. Issue 2.

22) Suguna N, Thanushkodi K, 2010, "An improved k-nearest neighbor classification using genetic algorithm". International journal of computer science issues. Vol 7. No. 4.

23) Rukundo Olivire. Hangqiang Cao, 2012, "References Nearest neighbor value interpolation". International journal of advanced computer science and applications. Vol 3. No. 4.

24) Mai Shouman. Turner Tim, Stocker Rob, 2012, "Applying k-nearest neighbor in diagnosis heart disease patients". International journal of information and education technology. Vol 2, No 3.

25) Moosavian A. Ahmadi H. Tabatabaeefar, 2013, "Comparison of two classifiers: k-nearest neighbor and artificial neural network, for fault diagnosis on a main engine". Shocks and vibrations. Vol 20.

26) Hassanat Ahmad Basheer, 2014, "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach". International journal of computer science and information security. Vol 12, No. 8.

27) Aggraval Rhasmi, 2016, "A modified K-nearest neighbor algorithm using feature optimization". International journal of engineering and technology. Vol 8, No 1.

28) Anchal, 2013, "A review of data classification using k-nearest neighbor

algorithm". International journal of emerging technology and advanced engineering. Vol 3. No. 6.

29) Chiang Tsung-Hsien, Hung-Yi Lo, Shou-De Lin, 2012, "A ranking based KNN approach for multi-label classification". Asian conference on machine learning JMLR; workshop and conference proceedings.81-96.

30) Fang Haw-Ren, Jie Chen, Saad Yousef, 2010, "Fast approximate kNN graph construction for high dimensional data via recursive lanczos bisection". Journal of machine learning research.

31) Bhatia Nitin, Vandana, 2010, "Survey of nearest neighbor techniques". International journal of computer science and information security. Vol 8. No 2.

32) Baur Brittanym Bozdag Serdar, 2016, "A feature selection algorithm to compute gene centric methylation from probe level methylation data". PLoS ONE. Vol 11, No. 2.

33) Cheng Xu, Hongzhu Qu, Guangyu Wang, Bingbing xie et al, 2015, "A novel strategy for forensic age prediction by DNA methylation and support vector regression model". Scientific report.

34) Rakyan VK, Beyan H, Down TA, Hawa MI et al, 2011, "Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis". PLoS Genet Sep;7(9):e1002300. (GSE56606)

35) Liu J, Morgan M, Hutchison K, Calhoun VD, 2010, "A study of the influence of sex on genome wide methylation". PLoS One Apr 6;5(4):e10028. (GSE34035).

36) Arner P, Sinha I, Thorell A, Rydén M et al, 2015, "The epigenetic signature of subcutaneous fat cells is linked to altered expression of genes implicated in lipid metabolism in obese women". Clin Epigenetics;7:93. (GSE24884).

37) Horvath S, Zhang Y, Langfelder P, Kahn RS et al, 2012, "Aging effects on DNA methylation modules in human brain and blood tissue". Genome Biol Oct 3;13(10):R97. (GSE24884).