DNA METHYLATION IN AGING: BEYOND THE LINEAR PROCESSES

O. Vershinina[1,*], *M. Ivanchenko*[1], *C. Franceschi*[1,2]

[1] Laboratory of Systems Medicine of Healthy Aging, Lobachevsky State University, Nizhny Novgorod, 603950, Russia
[2] University of Bologna and IRCCS Istituto delle Scienze Neurologiche di Bologna (ISNB), Bologna, 40139, Italy

Abstract. Methylation of DNA cytosine bases is a key epigenetic modification that plays an important role in the regulation of gene expression and the formation of the epigenome. Numerous studies of the human genome show that there is a close relationship between DNA methylation, age and sex of a person. Until now, the popular model has been the linear change in the methylation level with age. Here we find a fundamentally different DNA methylation behavior, namely the nonlinear dependence of the methylation level on age. We identify CpG probes whose methylation changes exponentially with age or according to a power law, and perform Gene Ontology enrichment analysis of the latter. Our results are relevant to understanding how DNA methylation changes with age and the found nonlinear CpG sites can be used to construct new epigenetic clocks.

Keywords: DNA methylation; aging; nonlinear epigenetic biomarkers.

List of abbreviations

DNA – Deoxyribonucleic acid
CpG – cytosine – phosphate – guanine
CH3 – methyl group
GO – Gene Ontology
GEO – Gene Expression Omnibus
NCBI – The National Center for Biotechnology Information
REVIGO – reduce and visualize Gene Ontology
BP – Biological Processes

Introduction

Aging is a complex process characterized by global physiological changes in the body. A distinctive feature of aging is epigenetic modifications of the genome that affect gene expression and regulation. The key epigenetic mechanism is DNA methylation, which involves the addition of a methyl CH3 group to the cytosine bases of DNA. Curiously, abundant experimental evidence suggests age-related changes to DNA methylation are associated with an increased risk of many diseases such as diabetes (Volkmar et al., 2012), Alzheimer's disease (De Jager et al., 2014) cardiovascular disease (Zhong et al., 2016) and cancer (Klutstein et al., 2016). There is also a search for epigenetic biomarkers of aging, the methylation of which is significantly different in young and older people (Bocklandt et al., 2011; Bell et al., 2012; Garagnani et al., 2012). Moreover, DNA methylation is used to quantify intrinsic ageing by developing various epigenetic clocks to predict biological age based on methylation data (Horvath, 2013; Horvath et al., 2016).

The overwhelming majority of studies on aging biomarkers have focused on detecting linear biomarkers, whose methylation level changes linearly with age. However, linear models do not always reflect the real dependence of DNA methylation level on age, and therefore, recently, attempts have been made to build nonlinear mathematical models of DNA methylation dynamics (Zagkos et al., 2019) and to use nonlinear models to predict biological age (Bekaert et al., 2015; Snir et al., 2019). Although earlier it has been discovered that methylation of CpG probes can change nonlinearly until adulthood (Horvath, 2013), specific CpG sites with methylation change law that differ from linear one during life were not reported yet. For this reason, here we have developed a procedure for finding nonlinear (power-law and exponentially) epigenetic biomarkers of aging and carried out GO pathway enrichment analyses of selected nonlinear probes.

Material and methods

Methylation Datasets. We consider two blood-based Illumina 450k datasets including only healthy subjects: GSE40279 (Hannum et al., 2013), GSE87571 (Johansson et al., 2013) from the Gene Expression Omnibus (GEO) datasets repository (Barrett et al., 2009) of The National Center for Biotechnology Information

(NCBI). The total number of subjects in the GSE40279 dataset is 656, of which 338 are females and 318 are males aged 19 and 101, and the total number of subjects in the GSE87571 dataset is 729, of which 388 are females and 341 are males between the ages of 14 and 94 years old. We chose these datasets as they have the largest age range among the currently available DNA methylation datasets that are critical for identification and distinction nonlinear and linear trends.

Data processing. Raw data files for GSE87571 were extracted and pre-processed using *minfi* and normalized using the *preprocessFunnorm* function from the Bioconductor package (Aryee et al., 2014). For GSE40279 dataset the analyses were carried out on pre-processed beta values available in GEO, but as the authors claim, GSE40279 beta values were adjusted for internal controls by the Illumina's Genome Studio software but not normalized. In addition, cross-reactive and polymorphic probes (Zhou et al., 2017) and probes on the X and Y chromosomes were excluded from further consideration. After performing the pre-processing steps, 414505 and 414950 probes remained for GSE40279 and GSE87571, respectively. Altogether 414505 probes were common to the two datasets and were included in further analysis. We also consider subsets of males and females separately, as methylation is known to be sex-specific (Liu et al., 2010; Tapp et al., 2013; Singmann et al., 2015; Yousefi et al., 2015; Yusipov et al., 2020).

Identification of CpG probes with nonlinear changes in methylation with age. We focus our attention only on CpG sites whose methylation levels change significantly with age. For this beta values for each probe were fitted to a linear regression model (through *OLS* function from *statsmodels* module for Python) using chronological age as covariables. We remove CpG site from consideration if its linear regression slope is less than 0.001 and its linear regression determination coefficient is less than 95% percentile for the distribution of determination coefficient of all sites.

To identify nonlinear CpG probes we fit beta values with a linear regression model for selected significantly age-associated sites in different scales: i) untransformed beta values from untransformed age (Fig. 1A), Eq. 1; ii) logarithmic beta values from logarithmic age (Fig. 1B), Eq. 2; iii) logarithmic beta values from untransformed age (Fig. 1C), Eq. 3,

$$\beta = s \cdot age + i, \quad (1)$$

$$\ln \beta = l \cdot \ln age + m, \quad (2)$$

$$\ln \beta = q \cdot age + p, \quad (3)$$

where β are beta values, s, i, l, m, q, p are fitting parameters.

Linear regression of beta values versus age in the logarithmic axes matches to the power law $\tilde{\beta} = e^m \cdot age^l$ in original axes; whereas regression in the semi-logarithmic axes matches to the exponential law $\tilde{\beta} = e^{q \cdot age + p}$ in original axes (see Fig. 1D).

The quality of models is evaluated by linear regression determination coefficient calculated for three resulting fits shown in Fig. 1D. This coefficient is calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \quad (4)$$

where $SS_{res} = \sum_{i=1}^{M} (\beta_i - \tilde{\beta}_i)^2$ is the residual sum of squares, $SS_{tot} = \sum_{i=1}^{M} (\beta_i - \langle\beta\rangle)^2$ is the total sum of squares, $\langle\beta\rangle = \frac{1}{M}\sum_{i=1}^{M} \beta_i$ is the mean of the observed data, $\beta_1, \beta_2, \ldots, \beta_M$ are beta values of specific CpG, $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_M$ are fitted beta values and M is the number of methylation points, that is, the number of subjects.

The coefficient of determination is a measurement, known as the «goodness of fit», is represented as a value between 0.0 and 1.0, where 1.0 indicates a perfect fit that is a highly reliable model. Moreover, we calculate power-law/exponential fits of complementary beta values, $1 - \beta$, on age and take them as better models if their determination coefficient is at least 5% higher than corresponding fits of beta values, β.

Finally, we compare three determination coefficients of linear, power-law and exponential fit and select the fit with the highest value of this characteristic as the best model reflecting the change in methylation levels with age.

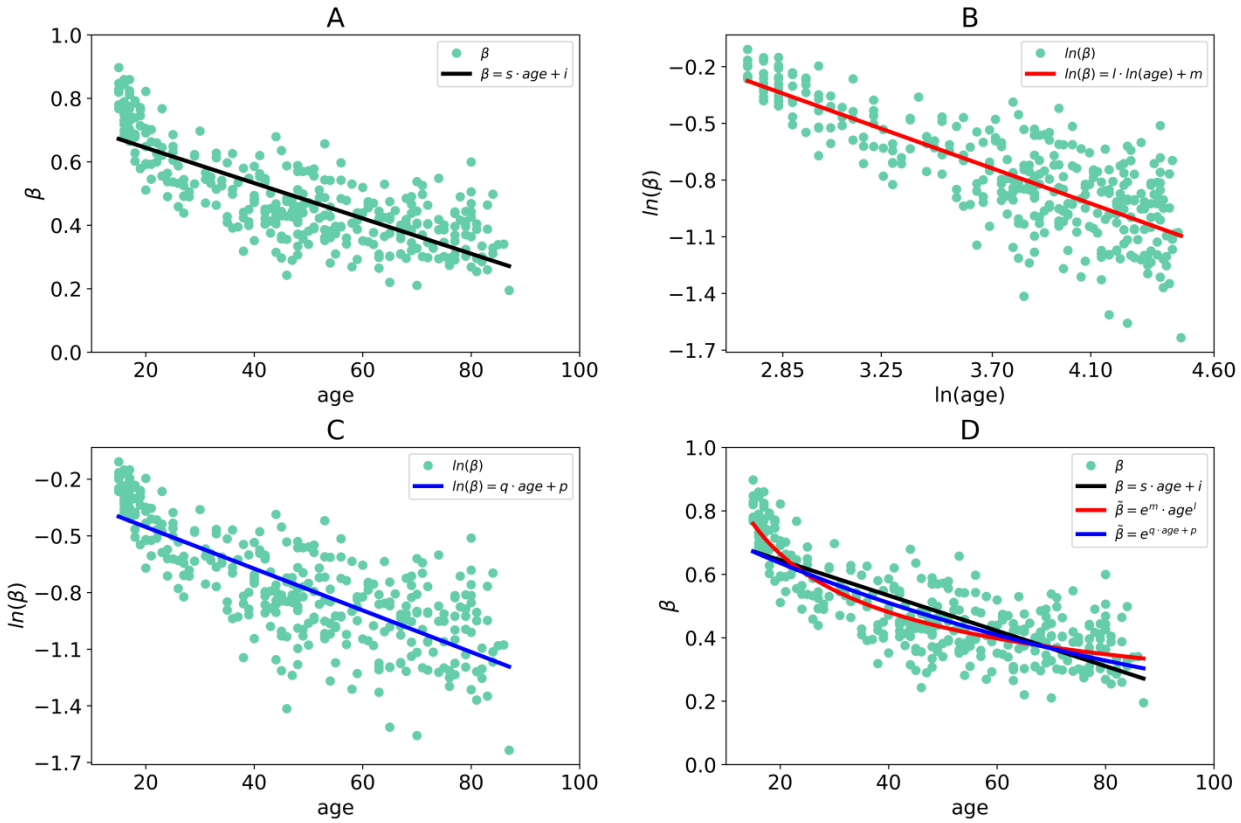

Fig. 1. Illustration of the algorithm for nonlinear CpG sites determining. A) linear regression of beta values versus age in the original axes. B) linear regression of beta values versus age in semi-logarithmic axes. C) linear regression of beta values versus age in logarithmic axes. D) the resulting nonlinear fits in the original axes

Gene Ontology Analysis. To take into account the fact that several CpG sites can correspond to the same gene, we perform Gene Ontology enrichment using the *methylglm* function implemented in the *methylGSA* Bioconductor package (Ren & Kuan, 2019) with default settings. This function allows carrying out the GO enrichment analysis of gene list after adjusting for the number of CpG sites. The list of GpGs with p-values of linear regression is used as input to the *methylglm* function. $P < 0.05$ is regarded as statistical significance. For more information on the ontology, we apply REVIGO (Supek et al., 2011) to GO term lists to remove redundant terms and to combine the remaining similar terms into even larger functional-semantic clusters.

Results

Identification of nonlinear probes. We selected CpG probes whose methylation is significantly related to age and limited ourselves to considering 6844/6954 CpG sites in the male/female subsets and 14244/12962 probes in the male/female subsets in GSE40279 and GSE87571 datasets, respectively. The number of probes satisfying the selection criteria was drastically lower in GSE40279 dataset compared to GSE87571. This difference can be attributed to the fact that only not normalized beta values were available in GEO for this dataset.

Among significantly age-associated probes, we identified CpGs with nonlinear dependence of the methylation level on age using the technique described in the Methods section. We considered CpG to be nonlinear if the nonlinear fit was 1% better than the linear one. The number of identified CpG sites having linear and nonlinear (power-law or exponential) age-associated methylation changes is presented in Table 1. Most nonlinear probes are probes whose methylation levels change according to the power law with aging. The number of power-

Table 1

The number of identified CpG probes having different laws (linear and nonlinear) of age-associated methylation changes for GSE40279 and GSE87571 datasets

	GSE40279	GSE87571
Males	**6844**	**14244**
Linear	5906	9780
Power law	934	4419
Exponential	4	45
Females	**6954**	**12962**
Linear	6638	10365
Power law	301	2564
Exponential	15	33

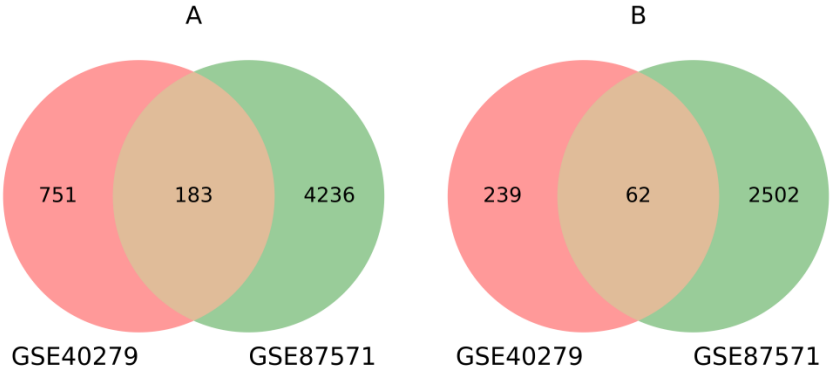

Fig. 2. Venn diagrams showing the intersection of nonlinear (power-law) CpG sites of GSE40279 and GSE87571 datasets; A) for males, B) for females

law sites ranges from 4 to 31% of the total number of sites considered, depending on gender and database. Intriguingly, the number of power-law CpGs is much higher in males than in females, so the nonlinearity of DNA methylation is sex-specific.

Next, we found common nonlinear (power-law) CpGs for GSE40279 and GSE87571 datasets, the number of which is shown in Fig.2, but we could not find common ones among the exponential probes.

Examples of typical CpG probes whose male's methylation level changes linearly, according to a power law and exponentially with age are presented in Fig. 3. CpG cg16867657 corresponding to ELOVL2 gene is shown as an example of a linear CpG site (see Fig. 3A, D). The dynamics of change in the methylation level of the power-law probes (see Fig. 3B, E) is characterized by a faster change of methylation in young people and slower in the elderly, in contrast with an average rate. Exponential changes in methylation with age show the opposite behavior. On the graphs of nonlinear sites in GSE40279 dataset, the difference between nonlinear and linear fits does not look as significant as in GSE87571 dataset. This is because methylation data in GSE40279 dataset is uneven and contains few subjects under the age of 40, which directly affects the behavior of nonlinear fits.

GO functional enrichment analysis

To explore the biological functions of the nonlinear CpGs, we performed GO pathway enrichment analyses. We looked at lists of power-law probes that were common to the two datasets and divided them into 3 groups: CpG sites

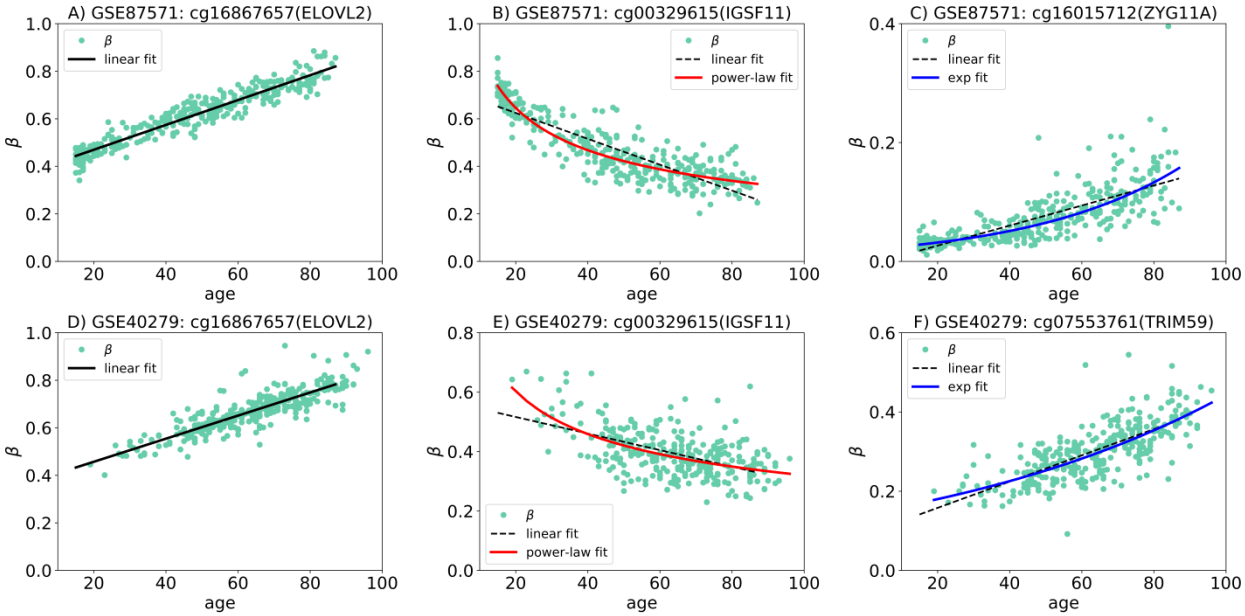

Fig. 3. Examples of CpG probes whose methylation level changes linearly, according to a power law, exponentially with age in GSE87571 dataset (A, B, C) and in GSE40279 dataset (D, E, F). The black dotted line indicates the linear fit for comparison

that are nonlinear in both males and females (17 CpGs); CpG sites that are nonlinear only in males (166 CpGs) and CpG sites that are nonlinear only in females (45 CpGs). We used REVIGO to remove redundant terms from GO term lists and concentrated on describing «Biological Processes» (BP).

Analysis of a group of sites, the methylation of which changes nonlinearly with age in both males and females, identified 5 most significant BP: GO:0007188 (adenylate cyclase-modulating G-protein coupled receptor signaling pathway), GO:0006835 (dicarboxylic acid transport), GO:0035249 (synaptic transmission, glutamatergic), GO:0007215 (glutamate receptor signaling pathway) and GO:0007187 (G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger). These BP are a subset of two major biological functions - adenylate cyclase-modulating G-protein coupled receptor signaling pathway and dicarboxylic acid transport.

For the group of sites, the methylation of which changes nonlinearly only in males, enrichment analyses revealed 4 most significant BP: GO:0031644 (regulation of neurological system process), GO:0098742 (cell-cell adhesion via plasma-membrane adhesion molecules), GO:0045185 (maintenance of protein location) and GO:0007156 (homophilic cell adhesion via plasma membrane adhesion molecules).

Finally, analysis of a group of sites, the methylation of which changes nonlinearly only in females, identified 2 most significant BP: GO:0030518 (intracellular steroid hormone receptor signaling pathway) and GO:0050848 (regulation of calcium-mediated signaling).

Discussion

We applied regression analysis for DNA methylation data from two datasets (GSE40279 and GSE87571) to identify CpG probes whose methylation levels change nonlinearly with age (according to the power law or exponentially). We found that a large number of sites have power-law methylation changes with age, which characterizes a rapid change in methylation at a young age and a slower change in the elderly. Interestingly, there were many more such sites for males than females. Thus, the nonlinearity of DNA methylation is sex-spe-

cific. We have also identified several dozen sites whose methylation changes exponentially.

Next, we selected a small number of power-law sites that are nonlinear in both GSE40279 and GSE87571 datasets and performed GO pathway enrichment analyzes. Pathway analysis of general nonlinear CpGs for males and females indicated enrichment in two global biological processes: adenylate cyclase-modulating G-protein coupled receptor signaling pathway and dicarboxylic acid transport. Considering probes that are nonlinear only in males, we got that the most enriched biological functions are regulation of neurological system process, maintenance of protein location and cell-cell and homophilic cell adhesion via plasma-membrane adhesion molecules. Analyzing the enrichment of CpG sites that are nonlinear only in females, we observed other significant biological processes: intracellular steroid hormone receptor signaling pathway and regulation of calcium-mediated signaling.

While conducting the research, we aimed to find reliable nonlinear DNA methylation changes not only in one specific dataset, so we considered two datasets. But at the same time, we faced some limitations: the analyzed datasets differ in terms of data preprocessing procedures and the uniformity of subject distribution by age. In particular, only not normalized beta values were available in GEO for GSE40279 dataset, and the data are unevenly distributed and contain few methylation values for young people. But despite the limitations that have arisen, we have identified for the first time specific nonlinear biomarkers of aging, the law of methylation changes of which differs from linear one. In the future, it is possible to build novel epigenetic clocks that will be based on nonlinear changes in DNA methylation level.

Acknowledgements

This work was supported by the grant of the Ministry of Education and Science of the Russian Federation Agreement No. 074-02-2018-330.

References

ARYEE M.J. et al. (2014): Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

BELL J.T. et al. (2012): Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PloS One*, **8**, e1002629.

BARRETT T. et al. (2009): NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, **37**, D885-890.

BOCKLANDT S. et al. (2011): Epigenetic predictor of age. *PloS One*, **6**, e14821.

BEKAERT B., KAMALANDUA A., ZAPICO S.C., VAN DE VOORDE W. & DECORTE R. (2015): Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics*, **10**, 922-930.

DE JAGER P.L. et al. (2014): Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature Neuroscience*, **17**, 1156–1163.

GARAGNANI P. et al. (2012): Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, **11**, 1132–1134.

HANNUM G. et al. (2013): Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, **49**, 359–367.

HORVATH S. (2013): DNA methylation age of human tissues and cell types. *Genome Biology*, **14**, R115.

HORVATH S. et al. (2016): An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology*, **17**.

JOHANSSON A., ENROTH S. & GYLLENSTEN U. (2013): Continuous aging of the human DNA methylome throughout the human lifespan. *PloS One*, **8**, e67378.

KLUTSTEIN M., NEJMAN D., GREENFIELD R. & CEDAR H. (2016): DNA methylation in cancer and aging. *Cancer Res.*, **76**, 3446–3550.

LIU J., MORGAN M., HUTCHISON K. & CALHOUN V.D. (2010): A study of the influence of sex on genome wide methylation. *PLoS One*, **5**, e10028.

REN X. & KUAN P.F. (2019): methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics*, **35**, 1958–1959.

SINGMANN P. et al. (2015): Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics & Chromatin*, **8**.

SNIR S., FARRELL C. & PELLEGRINI M. (2019): Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, **14**, 912-926.

SUPEK F., BOŠNJAK M., ŠKUNCA N. & ŠMUC T. (2011): REVIGO summarizes and visualizes long lists of Gene Ontology Terms. *PLoS One*, **6**, e21800.

TAPP H.S. et al. (2013): Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa. *Aging Cell*, **12**, 148–155.

VOLKMAR M. et al. (2012): DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO Journal*, **31**, 1405-1426.

YOUSEFI P., HUEN K., DAVE V., BARCELLOS L., ESKENAZI D. & HOLLAND N. (2015): Sex differences in DNA methylation assessed by 450 k BeadChip in newborns. *BMC Genomics*, **16**.

YUSIPOV I. et al. (2020): Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *BioRxiv*.

ZAGKOS L., AULEY M.M, ROBERTS J. & KAVALLARIS N.I. (2019): Mathematical models of DNA methylation dynamics: Implications for health and ageing. *Journal of Theoretical Biology*, **462**, 184–193.

ZHONG J., AGHA G. & BACCARELLI A.A. (2016). The role of DNA methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. *Circ. Res.*, **118**, 119–131.

ZHOU W., LAIRD P.W. & SHEN H. (2017): Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.*, **45**, e22.