Master Degree Project

UNIVERSITY OF SKÖVDE

DNA METHYLATION AGE ACCELERATION IN MULTIPLE SCLEROSIS

Master Degree Project in Bioinformatics:
Thesis report
Spring term 2018

Eleftheria Theodoropoulou (a17eleth@student.his.se)

Supervisor: Zelmina Lubovac (zelmina.lubovac@his.se)
Examiner: Björn Olsson (bjorn.olsson@his.se)
External Supervisors: Maja Jagodic (Maja.Jagodic@ki.se)
Francesco Marabita (francesco.marabita@ki.se)

To Tobias.

Contents

Abbreviations

AAR	Age acceleration residual
BMI	Body mass index
BMIQ	Beta-mixture quantile dilation
CNS	Central nervous system
EEAA	Extrinsic epigenetic age acceleration
DMF	Dimethyl fumarate (Tecfidera)
DNAmAge	DNA methylation age
FunNorm	Functional normalisation
Gran	Granulocytes
IEAA	Intrinsic epigenetic age acceleration
MS	Multiple sclerosis
NK	Natural killer cells
Noob	Normal-exponential using out-of-band probes
QN	Quantile normalisation
RTX	Rituximab
SQN	Subset quantile normalisation
SWAN	Subset-quantile within array normalisation

Abstract

Age acceleration is a measure indicating if a tissue is aging at an expected rate or not. In this study, the epigenetic clock was used to calculate age acceleration based on DNA methylation values in Multiple Sclerosis datasets. The samples were of whole blood, purified blood cell types and neurons and included individuals with the disease, as well as controls. Various factors were explored for their effect on the age acceleration in the context of the disease. In addition, three different normalisation options (no normalisation, Noob and Funnorm normalisation) were compared in order to assess their effect on the output of the epigenetic clock algorithm. Finally, a workflow was proposed for the epigenetic clock analysis, highlighting suitable methods for processing, analysing statistically and visualising the data.

1. Introduction

Epigenetics study the alterations in gene regulation and function that are heritable and do not involve changes in the gene sequence. These alterations in gene function may occur during a natural developmental process or can be induced by environmental factors. (Conerly and Grady, 2010)

DNA methylation

DNA methylation, the addition of a methyl group to a cytosine of a cytosine-phosphate-guanine (CpG) sequence, is an epigenetic control mechanism, pivotal for the functional regulation of the genome of vertebrates. As it is suggested, it is involved in various genomic processes, e.g. gene transcription and expression regulation, gene silencing, and chromosomal stability (Rakyan *et al.*, 2004). Aberrant DNA methylation can lead to hyper- or hypo-methylated positions or regions of DNA, which can result in various implications and ultimately in the development of a disease, such as cancer (Conerly and Grady, 2010) and autoimmune diseases (Richardson, 2003). In addition, it is known that DNA methylation changes as the age of the cell, tissue or organism progresses (Marioni *et al.*, 2015), and in particular two studies developed the prediction of chronological age in humans using DNA methylation values (Hannum *et al.*, 2013; Horvath, 2013).

Currently, the state of the art array for studying DNA methylation of the human DNA is Illumina's Infinium MethylationEPIC BeadChip. It contains >850000 methylation sites (CpGs), covering >99% known genes (*MethylationEPIC BeadChip by Illumina*, 2017). Its predecessor, Infinium HumanMethylation450 Bead Chip array by Illumina, has 485000 individual CpGs, and >90% of these are shared in the EPIC array. As of 2016 there are more than 360 publications that used Illumina methylation arrays (Kurdyukov and Bullock, 2016).

Methylation values, β (beta), are between zero and one and represent the percentage of methylation of a CpG position across the cell population in a sample.

Age acceleration

The term age acceleration refers to the residual value of a linear regression model that is estimated between the chronological age and the DNA methylation (predicted) age of a cell/tissue (sample). This age acceleration can be a result of various factors that affect DNA methylation changes; these factors involve normal developmental processes, environmental factors, disease state and aging. In order to derive the predicted age of a sample, an "epigenetic clock" is used. The term epigenetic

clock refers to the age predictor model development that allows the estimation of the age of cells or tissues based on their DNA methylation status (Horvath, 2013).

Various epigenetic age estimators have been developed to assess the biological age (biological state) of individuals, with two being highly accurate (Horvath and Raj, 2018). One of them was developed to calculate age from whole blood samples, is dependent on blood cell count changes over time and measures the extrinsic epigenetic age acceleration (EEAA), that is connected to the aging of the immune system and the gradual loss of its protective role (Hannum *et al.*, 2013). This epigenetic age predictor can only be used on whole blood samples and is affected by various environmental factors (Quach *et al.*, 2017). The other is a multi-tissue predictor of age, independent of changes in cell counts over time and thus measures the intrinsic epigenetic age acceleration (IEAA), which is independent of cell type (Horvath, 2013). The extrinsic model is based on 71 CpG markers, while the intrinsic has 353 CpG markers of age. Both models provide high correlation values between the DNA methylation age and the chronological age across a multitude of tissues for the Horvath clock, and whole blood for the Hannum clock, which is indicator of their high accuracy as age predictors (Hannum *et al.*, 2013; Horvath, 2013). Both age acceleration measures are publicly accessible online via the epigenetic clock tool created by Steve Horvath (*DNA methylation age and the epigenetic clock*, 2013a) and the software is based on his published epigenetic clock with 353 CpG sites as epigenetic markers of age (penalized linear regression model, elastic net) (Horvath, 2013). This software provides among other calculations, an output on the age acceleration residual (AAR) (from the linear regression model – for all tissues) and both the intrinsic and extrinsic epigenetic age acceleration measures (for whole blood).

Other developed age predictors have low accuracy or are not applicable across a multitude of tissues (Horvath and Raj, 2018). In addition, a new epigenetic age estimator, DNAm PhenoAge (Levine *et al.*, 2018) has been published recently, and it provides an estimate for morbidity and mortality similarly to clinical phenotypic values measured in blood; e.g. Insulin, Glucose, Triglyceride, blood pressure etc. This age estimator is closer to the Hannum clock, since they are both developed using whole blood data (Horvath and Raj, 2018).

Multiple Sclerosis

Multiple sclerosis (MS) is a serious neurological condition; a progressive disease that occurs unpredictably and can have milder symptoms like fatigue and depression, to grave symptoms such as severe mobility problems and blindness. It is considered an immune mediated disease, where myelin, the nerve insulating substance that is responsible for the proper functioning of the nervous system is attacked by the immune system of oneself (*National MS Society*).

Multiple sclerosis is mostly diagnosed between the ages of 20 and 40, but the disease onset might be earlier. According to the “MS international federation” there are more than 2,300,000 people around the world that have been diagnosed with MS (*Multiple Sclerosis International Federation*, 2016). In 2015, there were reported to be 17500 people living with MS in Sweden (*European Multiple Sclerosis Platform*, 2015). Currently, there is no cure for MS; however, there are several drugs (e.g. RTX – Rituximab and DMF - Dimethyl fumarate with the commercial name Tecfidera) that deal with the symptoms of the disease to make it more manageable.

There is some genetic susceptibility to the disease; e.g. individuals that carry the HLA-DRB1*15:01 risk variant and do not carry the HLA-A2 protective variant (Sawcer *et al.*, 2011). Although the

possible triggers for the disease are considered to be environmental and are currently unknown, there are indications that high latitude, smoking, low vitamin D levels and being infected by Epstein Barr Virus in particular are potential triggers (*Multiple Sclerosis International Federation*, 2016; Wergeland *et al.*, 2016). In addition, having high body mass index (BMI>27) at the age of 20 is a risk factor for the disease (Hedström, Olsson and Alfredsson, 2012).

Design and Aim of the Study

The process of aging itself has been known to affect the efficiency of the immune system, making the organism susceptible to various diseases that rely on the appropriate response of the immune system, such as cancer and auto-immune diseases (Castelo-Branco and Soveral, 2014). In addition, it has been reported that the epigenetic clock can reveal information about the biological age, and age acceleration can provide independent predictions of all-cause mortality later in life (Marioni *et al.*, 2015; Perna *et al.*, 2016; Stölzel *et al.*, 2017). This means that the epigenetic clock is a useful tool in assessing whether an organism is aging in a healthy way or not.

Furthermore, since the development of the Horvath epigenetic clock in 2013, where several datasets were tested and found to have age acceleration (mostly cancer tissues), studies have been conducted that show age acceleration in the context of other health disorders too; obesity and the aging of the human liver (Horvath *et al.*, 2014), age acceleration in Down syndrome (Horvath, Garagnani, *et al.*, 2015), Parkinson's disease (Horvath and Ritz, 2015), coronary heart disease (Horvath *et al.*, 2016) and HIV-1 infection (Horvath and Levine, 2015).

In this study, the bioinformatics contribution was highly intertwined with the biological significance. A similar study had not yet been conducted in the context of MS, so the information acquired by the analysis of the results provides significant input on the relationship between age acceleration and the disease.

The first scientific aim of this study was on the bioinformatics contribution. The aim was to standardise pre-processing methods for preparing the input matrices for the epigenetic clock analysis and assess if there is an effect of the pre-processing and normalisation methods in the analysis. Due to the high number of normalisation methods available and applicable to DNA methylation data, and the fact that each normalisation method works slightly different on the raw data, a suitable method for the epigenetic clock analysis should be used in order to make current and future analyses comparable. Previous studies using the epigenetic clock algorithm do not mention the normalisation method used (Horvath *et al.*, 2014; Horvath and Ritz, 2015; Horvath, Garagnani, *et al.*, 2015; Lu *et al.*, 2017). Only a few articles state that they used the modified beta-mixture quantile method (BMIQ), originally developed by A. Teschendorff (Teschendorff *et al.*, 2013) and modified by Horvath (see "Materials and Methods" section) (Horvath and Levine, 2015; Horvath, Mah, *et al.*, 2015; Knight *et al.*, 2016) or dasen method (see "Methods not used") (Stölzel *et al.*, 2017). Furthermore, various statistical and visualisation analysis tools were used and discussed, and a workflow was proposed for analysing the output of the epigenetic clock algorithm. Ultimately, the bioinformatics contribution lies in the standardisation of the epigenetic clock analysis, from the pre-processing to the conclusions, so that future projects can benefit from an efficient design of the process and targeted analysis of the epigenetic clock output to identify significant results.

The second aim, with focus on biology, was to estimate the potential age acceleration of the samples of various datasets in order to compare the two epigenetic age acceleration measures in the context

of MS. Since MS is an immune-mediated disease, where immune cells are activated and cause inflammation in the central nervous system (CNS), there are two groups of cells that are involved in the disease; blood cells and brain cells. Therefore, compared to other diseases mentioned above, in this study the importance lies in how age acceleration estimates different between the "attacking" cells versus the "vulnerable" cells in MS. The various datasets provided different contrast settings in which the epigenetic clock analysis could be applied; e.g. disease status, cell and tissue type and therapy. As mentioned above, the two epigenetic age acceleration measures can provide information on different aspects (intrinsic/extrinsic) of the age acceleration of a sample based on its DNA methylation status, and therefore yielded different results based on the design of the experiment (contrasts and tissue types). Ultimately, the scientific input gained from this objective will lead to future projects to investigate the findings and expand on the role of MS in age acceleration of blood and brain cells.

2. Materials and Methods

In order to provide the reader with a clear picture of the process steps in this project, a brief workflow description is provided in Table 1.

Table 1. Workflow steps for complete epigenetic clock analysis of each dataset. The table includes the step order, main objective of each step, details on the implementation of each step and the Bioinformatic tool (program) in which each step will be realized.

Order	*Objective*	*Details*	*Program*
1	Sample annotation	Input required for epigenetic clock. Per dataset (batch)	R
2	Raw data reading	Per dataset separately	R – *minfi*
3	Pre-processing/normalisation	• None • Noob • Functional normalisation	R – *minfi*
4	β values matrix	Input required for epigenetic clock	R – *minfi*
5	Epigenetic clock analysis	Age acceleration residual calculation and two additional adjusted measures (for whole blood only)	Epigenetic clock algorithm (Horvath)
6	Analysis of results	Stratification according to biological/phenotypical variables	R

Data

Nine datasets of DNA methylation data obtained with Infinium HumanMethylation450 (450k) Bead Chip or HumanMethylationEPIC (EPIC) arrays by Illumina were used in this project. The datasets have different phenotype variables available for analysis, e.g. disease status, gender, therapy, and are whole blood samples or purified cell types (blood cells, brain cells). In total there are 706 samples

among the datasets, although the size of each dataset differs. It is worthy to mention that some datasets contained technical replicates, and some had multiple samples from the same individual; e.g. samples before, during and after treatment, or different purified cell types from the same individual at the same sampling date. This means that some chronological ages might be represented more than others, especially in smaller datasets. Table 2 provides more information on the datasets.

Table 2. This table contains several pieces of information for each dataset. The dataset name, as well as the batch number (the datasets were separated based on their batch, designating which samples were ran together). In addition, in the Sample column there is information on the number of samples (left) and number of individuals (right). Moreover, the tissue or cell type that comprises the dataset are given. Lastly, the variables describing the different contrasts are given.

Dataset Name	*Batch*	*Samples/Individuals*	*Tissue/Cell type*	*Contrasts/Variables*
Broad	B01	279/279	Whole blood	Various phenotypic and genotypic variables
Selected	B02	52/49	Whole blood	Various phenotypic and genotypic variables
CD14	B02	43/36	Monocytes (CD14)	Disease, gender
CD4_4CT	B03	36/33	CD4 T-cells	Disease, gender
CD8_CD19_4CT	B04	58/39	CD8 T-cells and B-cells (CD19)	Disease, gender, cell type
RTX	B05	70/17	CD4 and CD14	Disease, cell type, treatment
DMF	B06	96/31	CD4 and CD14	Disease, gender, cell type, treatment
Brain_pch	B07	12(-1)/2	Sorted neurons and bulk brain tissue	Disease, gender, brain related variables
Brain_ch1	B08	12/12	Sorted neurons	Disease, gender, brain related variables
Brain_ch2	B09	24/21	Sorted neurons and bulk brain tissue	Disease, gender, brain related variables

Each sample was carefully annotated in R. Age, gender, tissue (or cell) type are required by the epigenetic clock algorithm and subsequent analysis of the results (optional input, a comma delimited .csv file). Other phenotype and biological data were included in order to assess the biological relevance of the results at the end of the analysis (Step 6 in the workflow, Table 1). To this end, each sample was annotated according to the information recorded for the experimental design which produced the dataset. This means that not all datasets have the same variables describing phenotype and other biological data (Table 2).

The raw data (DNA methylation intensities – IDAT files) were read in R using the package *minfi* (Aryee *et al.*, 2014). The *minfi* package for R contains several functions that facilitate the loading, pre-processing, normalising and analysing DNA methylation data, and was therefore used in multiple steps of the project. The object created this way is complex and contains the raw data as well as phenotype and other biological data that correspond to each sample. The object created for each dataset was saved for further analysis.

Pre-processing and Normalisation

After loading the data, sample quality control (QC) analysis was conducted as part of the pre-processing of the data. The QC was conducted using the package *shinyMethyl* in R. Potential outliers were investigated further and pdf files with detailed QC data were produced using the *minfi* package. The two packages produce similar graphs for QC analysis. Only in the dataset "Brain_pch" was an outlier identified and removed before further analysis (final sample count for this dataset: 11). In

addition, three samples were removed from the DMF dataset, where cells in the sample were not pure enough to be considered purified for the specific cell type.

Table 3 provides information on the normalisation methods used in this project and what they correct for. The datasets were loaded and normalized separately using the methods Noob (Normal-exponential using out-of-band probes) (Triche *et al.*, 2013) and Funnorm (Functional normalisation) (Fortin *et al.*, 2014). Both methods are available in *minfi* package in R. The normalisation steps are depicted in Figure 1 below.

Functional normalisation (Fortin et al., 2014) has been tested and shown to perform robustly and efficiently compared to other normalisation methods in both 450k as well as the new EPIC arrays for DNA methylation (Fortin et al., 2014; Fortin, Triche and Hansen, 2017). The method is an unsupervised approach of quantile normalisation, using the control probes on the arrays to normalise the data and has been tested in studies where global methylation levels are expected to differ significantly. This makes the method of great interest to study in this project, where datasets have so different contrast levels. It is important to note that the R function that conducts functional normalisation also corrects for dye bias and background intensity by applying the Noob normalisation as a first step by default. Noob normalisation is a basic normalisation method that is typically used in combination with some quantile normalisation step. It performs background and dye bias correction by measuring non-specific fluorescence in opposite colour channel using type I probe design (Triche *et al.*, 2013).

It is important to note that even though BMIQ (Teschendorff *et al.*, 2013) is not part of the normalisation methods comparison for the objectives of this project, it was used as part of the epigenetic clock analysis and was implemented by the online algorithm in order to make the datasets comparable to the training set of the epigenetic clock (option selection in the data submission of the online tool). The version of BMIQ used by the online tool is slightly modified to fit a "golden standard" based on the training sets used for developing the epigenetic clock algorithm (Knight *et al.*, 2016).

Table 3. Normalization methods used for the epigenetic clock analysis. The table includes the method name, the main objectives and some relevant details, the type of normalization (within or between-array), and the type of data normalized (raw intensities of β values) with each method.

Method	*Objectives*	*Details*	*Normalization*	*Data normalized*
Noob	Measure non-specific fluorescence in opposite colour channel using type I probe design.	Background correction and dye bias.	Within-array	Raw intensities
Functional normalisation (FunNorm)	Remove technical variation using QN and control probes. Adjust separately for type II and type I probes.	No assumptions needed (unsupervised). Noob as first step.	Between-array	Raw intensities
Beta-mixture quantile dilation (BMIQ)	Adjust type II to type I probe distribution. Corrects probe design bias.	Done by epigenetic clock online algorithm (modified).	Within-array	β values

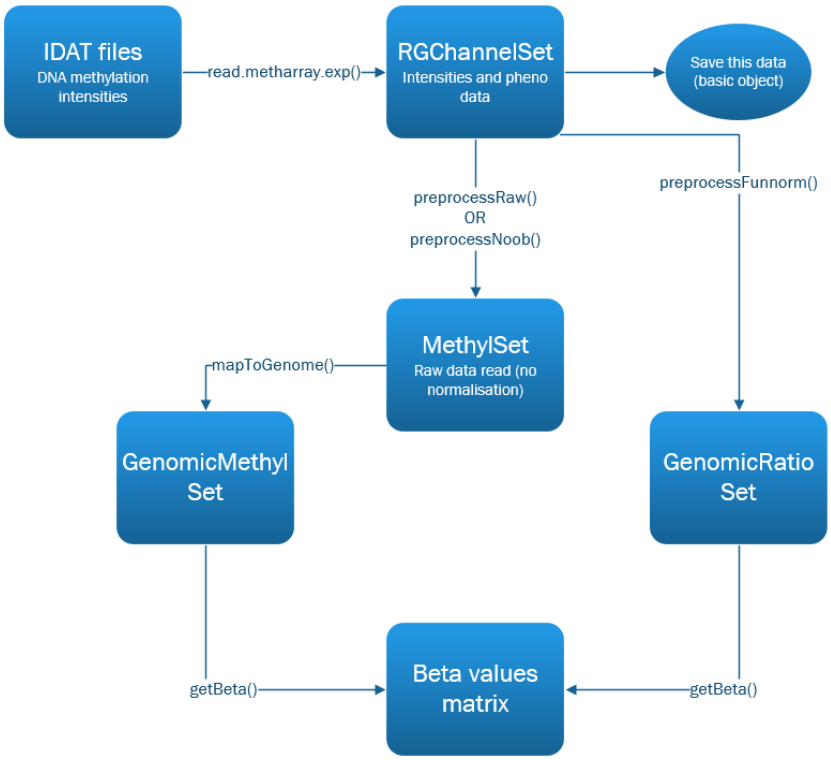


Figure 1. Steps involved in obtaining the beta values matrix with or without normalisation. The boxes show the data and R objects generated, while the functions used are on the arrows. The final product, regardless the path, is a beta values matrix, comprising the methylation values for each probe. All functions are included in *minfi* package, R.

Epigenetic Clock

From the object created in Step 2, Raw data reading (Table 1), using package *minfi* in R, the β values can be calculated, either from the raw data or after normalization using *minfi*, as shown in Figure 1. The final matrix (comma delimited .csv file) was created for each dataset and normalization method after extracting the probes required by the online algorithm (around 28.000 – provided by Horvath on the epigenetic algorithm site) including the 353 CpGs that are used in the model of age acceleration. The files were used as the mandatory input to the epigenetic clock algorithm.

The epigenetic clock algorithm (*DNA methylation age and the epigenetic clock*, 2013a) was used to calculate age acceleration in the samples of the aforementioned datasets. Normalisation option was selected for all datasets (and all normalisations), and advanced analysis for blood data was selected only for the datasets (batches) that contained such samples.

The output created by the algorithm includes various information; DNA methylation age, age acceleration residual (AAR) and the two other measures of epigenetic age acceleration (only after advanced analysis for whole blood data), intrinsic epigenetic age acceleration, IEAA (Horvath, 2013), and extrinsic epigenetic age acceleration, EEAA (Hannum *et al.*, 2013), are among the results. In addition, the algorithm returns information on missing beta values, probability of match to various tissue types, gender prediction, and other variables. By examining those results, it was possible to confirm the purity of the samples and avoid mismatches.

Statistical analysis and Visualisation

Different contrast settings of datasets as well as biological variables (tissue type, age, gender, genetic variants, and other MS and/or aging associated variables) were obtained for each dataset and used to investigate the two epigenetic age acceleration measures as well as the age acceleration residual between the different contrast levels.

Several steps of statistical analysis and visualization of the results were employed. Table 4 contains information on the methods used. R was used for implementation of all methods described.

Table 4. Statistical and visualization methods (plots) used to analyze the results of the epigenetic clock analysis.

Method	*Objective*
Scatter plot	Visualize the correlation between two variables and the distribution of samples from different contrast levels (e.g. Male VS Female etc.).
Correlation test	Estimate the correlation between two variables and fit of the model/method. (Pearson's correlation).
Error estimation	Estimate error of the model/method. (Median absolute difference).
Box plot	Demonstrate the distribution of the samples and visualise differences between groups using one or more factors.
Friedman's test	Non-parametric test to assess the significance of the observed difference between groups.
Bar plot	Visualize differences between different contrast levels of a variable among the samples.
T-test	Estimate the significance of the differences between two groups in a variable among the samples.
Linear model	Estimate the significance of the differences between two or more groups in one or multiple variables among the samples.
Linear mixed model	Include random effects to the linear model (e.g. effect of individual).

3. Results and Analysis

Relation to the aim

The aim of this project was extending in both bioinformatics and biology. A suitable combination of methods needed to be established for a better use of the epigenetic clock, to retrieve the potential biological information from this analysis. To this end, three normalisation options were considered, no normalisation (raw data), Noob and Funnorm normalisation. The reasoning behind this, is that Noob and Funnorm normalisation have been widely used in other DNA methylation studies and are known to perform well (Liu and Siegmund, 2016), and Noob normalisation is incorporated in Funnorm as a first step (Fortin *et al.*, 2014). Thus, the data is "gradually normalised" and the "extra" steps could be assessed for their usefulness in the context of epigenetic clock analysis. Analysis for the normalisation options and selection of the best one based on the reduction of technical variation in the datasets (reason why normalisation methods are developed) as well as the data fit to the predictor model (epigenetic clock), is described under "Normalisation options comparison".

In addition, an epigenetic clock analysis has not yet been done in the context of MS, and thus this is a novel exploration of age acceleration in MS patients. Especially since MS is an immune-mediated disease, by investigating the cells that "attack" (blood cells) as well as the cells that are "being attacked" (neurons), it was interesting to see how age acceleration differs in these two categories of cells compared to the controls. Additionally, the available datasets used in this project had different contrast levels and phenotypic variables, thus more than one biologically relevant question could be posed, and the right dataset could be used to explore a certain hypothesis. These contrasts involved among Gender and Disease status, receiving therapy for the disease (RTX or DMF), having a genetic risk for the disease (HLA risk allele), vitamin D levels, smoking status and body mass index (BMI), all associated with developing the disease, and lastly different cell type samples, even some belonging to the same individual. Hypotheses formed while observing the preliminary data shown in "First results exploration and hypothesis formulation") and further analysis was conducted to address those hypotheses, which is described in section "Linear models to answer biological questions".

Moreover, an efficient workflow proposition was made after discussion of the statistical methods used in this project and their pitfalls, in order to facilitate the analysis in future projects using the epigenetic clock algorithm (section "Discussion of methods and Workflow proposition").

Lastly, methods that were considered but were ultimately not used in this project are briefly discussed in section "Methods not used".

Normalisation options comparison

Firstly, the normalisation options were assessed using the technical replicates, and the predictor model fit based on correlation and error estimates.

- ### Technical replicates

Differences between the technical replicates of a batch is a measure of the size of technical variation in that batch. In order to evaluate the reduction in variability, the absolute difference in the calculated epigenetic clock variables between technical replicates was used. Because a technical replicate is defined as repeated measurement from the same biological material (DNA sample), it is expected that any observed difference can be considered a source of technical variation introduced

during the experimental and processing pipeline (Marabita *et al.*, 2013). Normalising the data serves the purpose of reducing technical variation, so that the biological variation can be highlighted.

Figure 2 depicts the differences between the technical replicates among the three normalisation options (raw data=no normalisation, Noob and Funnorm normalisation) for three out of nine datasets used in this project (replicate pairs ≥3). All the differences shown between the three normalisation methods are significant at the 0.05 level, apart from the Extrinsic Epigenetic Age Acceleration - EEAA measure in the Selected dataset. However, it is noteworthy that the replication is only adequate in the CD14 dataset, which had more replicate pairs (8 pairs) with respect to the other two datasets (3 pairs).

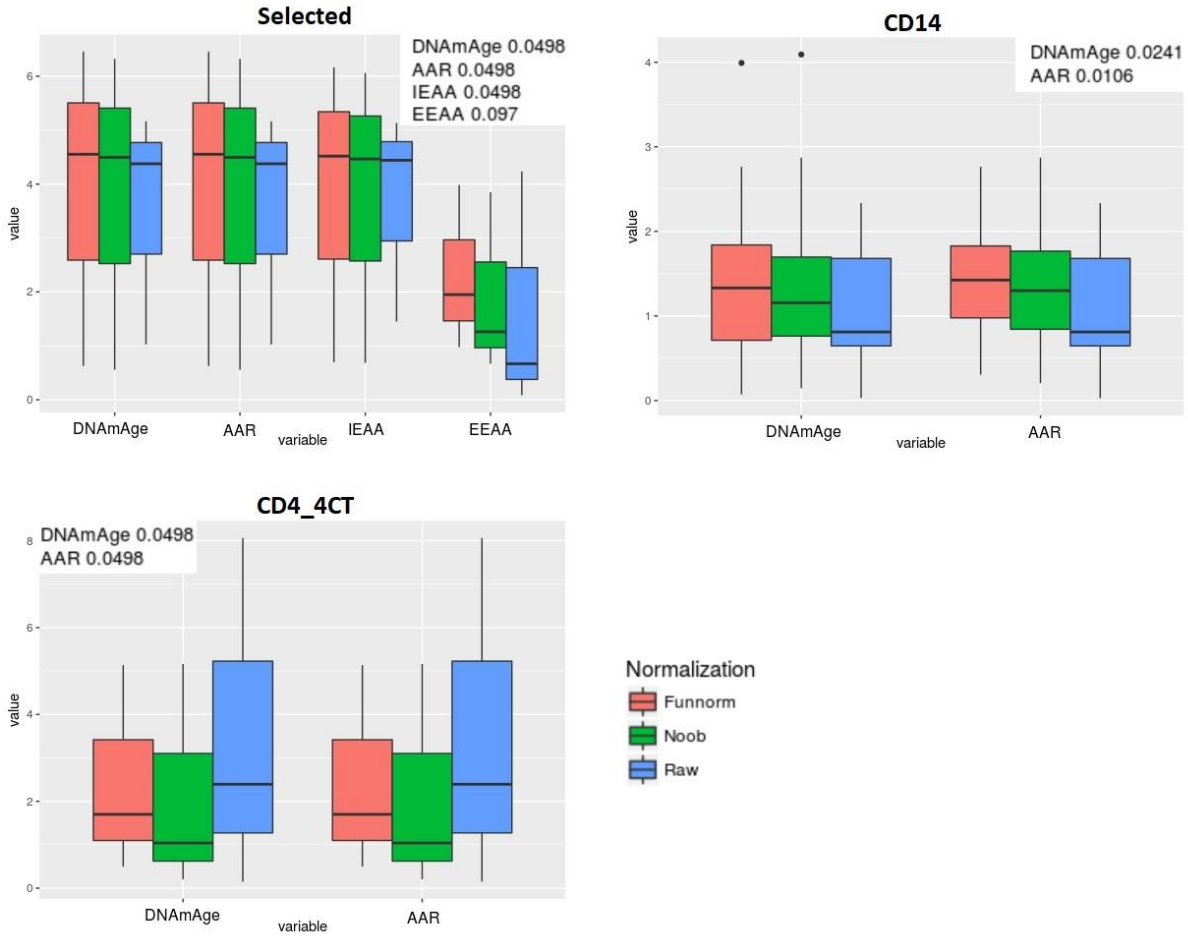


Figure 2. Box plots for Selected, CD14 and CD4_4CT datasets showing the distributions of the absolute differences between technical replicates (y-axis, value) for DNAmAge, AAR, IEAA and EEAA (where applicable). In red is Funnorm normalised data, in green is Noob normalised data and in blue is raw data. The p-values of the Friedman's test for each variable is found in the legends at the top of the graphs.

As seen in Figure 2, the data suggest that normalisation did not reduce the technical variation between the replicates compared to raw data in two out of three datasets (CD14 and Selected) for all measures of age acceleration (Age Acceleration Residual – AAR, Intrinsic Epigenetic Age Acceleration – IEAA and Extrinsic Epigenetic Age Acceleration – EEAA), as well as at the prediction of biological age (DNA methylation age – DNAmAge).

However, Noob normalisation was found to have the lowest technical variation in CD4_4CT. Although it is not conclusive which option performs better, the Noob and the Raw method were preliminarily selected.

- **Correlation values and error estimates**

The correlation and error estimation for DNA methylation age versus chronological age were checked for all datasets and all normalisations. This shows how well the predictor model (epigenetic clock algorithm) has worked for each dataset and provides another way to compare the normalization options. It shows how normalised data compare to the raw values. This was done by applying a Pearson's correlation test and estimating the error of the predictions (DNA methylation age) in years. The error is estimated by calculating the median absolute difference (|x-y|) between the two variables (DNA methylation age and chronological age). This measure is found in Horvath's tutorial for the epigenetic clock algorithm. Table 5 comprises information on these values for each dataset.

Out of the nine datasets (batches), only in batch 02 (B02 – Selected and CD14) was the raw data more significantly correlated between the chronological and the biological age. In all other datasets, Noob or Funnorm normalised data showed a slight increase in correlation estimate and a decrease in the error estimate, making the predictor model perform better with this data. Both normalisation methods seem to perform very similarly, so Funnorm does not seem to enhance the effect seen with Noob normalisation. Considered that Funnorm is a more complicated method and does not perform better in the technical variation reduction (see above), it does not add value to the analysis. Further analysis of the datasets was performed using only the Noob normalised data.

Table 5. Correlation, p-values and error estimates for all datasets/normalisations for the graphs DNA methylation age versus chronological age (fit of the model – age predictor). Under the Samples column, the number of samples (left) and number of individuals (right) are shown for each dataset (the batch number is shown in parentheses to indicate which datasets were processed together). The "best" combinations of Correlation, P-value and Error are shown in blue. No selection was made for brain_ch1, since p-values are not significant.

Dataset	*Samples*	*Normalisation*	*Correlation*	*P-value*	*Error (years)*
Broad (B01)	279/279	Raw	0.93	3.8e-126	5.3
		Noob	0.94	9e-130	3.3
		Funnorm	0.94	3.1e-130	3.3
Selected (B02)	52/49	Raw	0.87	1.2e-16	2.7
		Noob	0.87	3.8e-17	2.9
		Funnorm	0.87	6.2e-17	2.8
CD14 (B02)	43/36	Raw	0.93	5e-19	2.7
		Noob	0.93	5.1e-19	3.1
		Funnorm	0.93	7.3e-19	3.1
CD4_4CT (B03)	36/33	Raw	0.84	1.6e-10	4.5
		Noob	0.86	2.1e-11	2.5
		Funnorm	0.86	2.8e-11	2.7
CD8_CD19_4CT (B04)	58/39	Raw	0.78	7.5e-13	4.2
		Noob	0.77	1.3e-12	3.9
		Funnorm	0.77	1.2e-12	3.8
RTX (B05)	70/17	Raw	0.76	1.3e-14	4.2
		Noob	0.79	5.3e-16	3.5
		Funnorm	0.79	3.7e-16	3.5
DMF (B06)	96/31	Raw	0.61	5.7e-11	5.2
		Noob	0.61	3.8e-11	5.1
		Funnorm	0.63	7.3e-12	5.3
Brain_pch (B07)	11/2	Raw	0.91	0.00013	6.8
		Noob	0.91	8.4e-05	5.4
		Funnorm	0.93	3e-05	4.8

Brain_ch1 (B08)	12/12	Raw	0.57	0.054	8.3
		Noob	0.53	0.074	9.6
		Funnorm	0.56	0.056	8.2
Brain_ch2 (B09)	24/21	Raw	0.85	1.2e-07	17
		Noob	0.88	1.3e-08	11
		Funnorm	0.89	5.6e-09	11

First results exploration and hypothesis formulation

- **Algorithm prediction results: Scatter plots, correlation tests and error estimates**

In addition to Table 5, a scatter plot was created for each dataset and normalization option and a linear model line was added, in order to visualize the fit of the data (correlation between DNAmAge and chronological age). Figure 3 (graphs a, d, and g) shows the scatter plots made for the raw data and the two normalisations of the Broad dataset (B01). The Broad dataset was selected since it has the highest number of samples (279) among the datasets used in this study, it has no replicates or multiple samples from the same individual, and it is only made up of one tissue type (whole blood). To avoid showing too many graphs, only the correlation test and error estimates (statistical values) are shown for the rest of the datasets in Table 5.

Moreover, the chronological age was plotted against the age acceleration residual for all datasets and normalizations to confirm that the effect of age has been regressed out and there is no correlation. This is shown in Figure 3 (graphs b, e and h) as well. Lastly, the DNA methylation age was plotted against the age acceleration residual (Figure 2, graphs c, f and i), showing a slight correlation, but high error. This is expected, since age acceleration residual is deriving from DNA methylation values and is therefore connected to DNA methylation age (only chronological age is regressed out). These two types of scatter plots showed the same results for all other datasets, and therefore will not be shown.

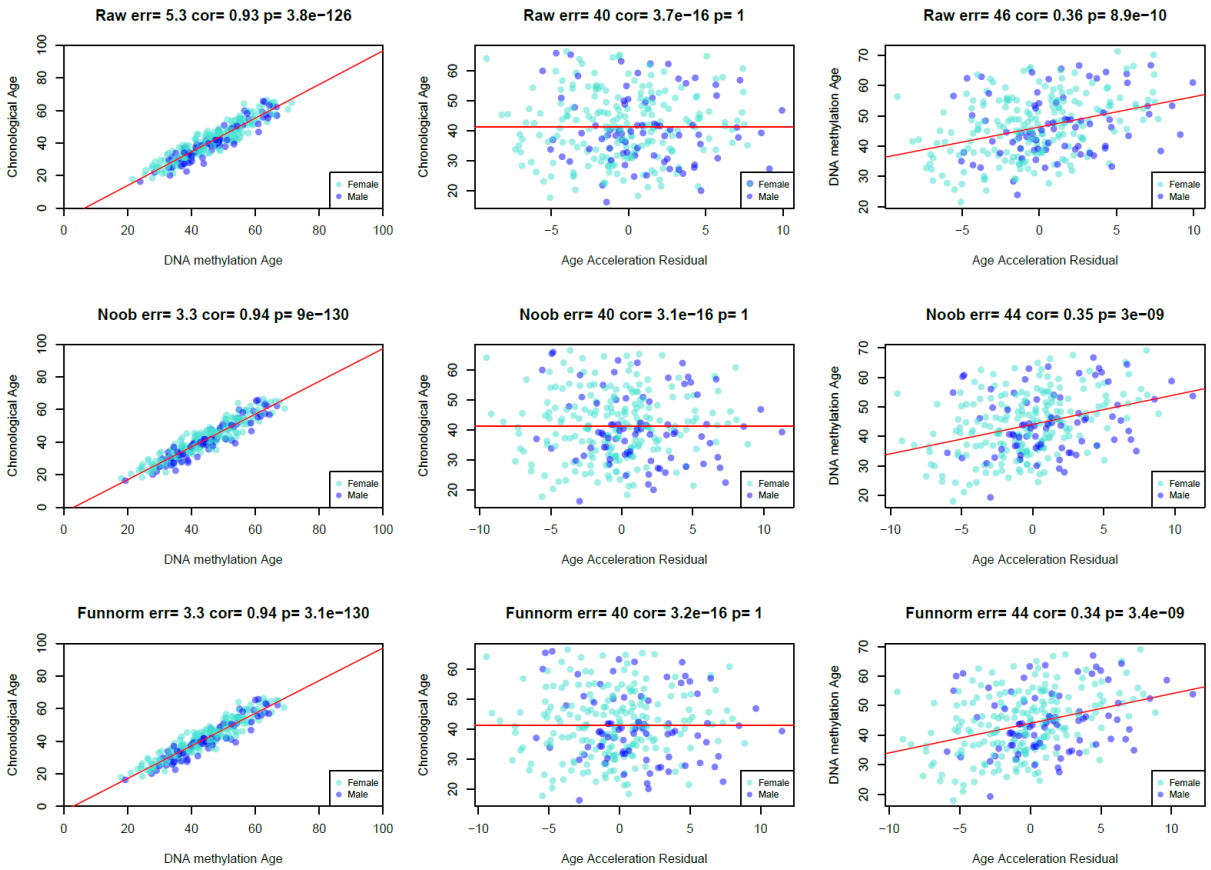


Figure 3. Scatterplots showing correlation between different age variables in the Broad dataset. a-c) Raw data, d-f) Noob normalized data and g-i) Funnorm normalized data. Light blue: Female samples, Dark blue: Male samples.

It is worthy to mention, that in this dataset there cannot be seen a higher variation with higher age, since the age upper limit is around 65-70 years, and it would have been more visible at higher ages (>80) (Figure 3 a, d and g).

Additionally, different colours were used to show different contrast groups in the data scatterplots; Male/Female, MS/Controls and different cell types, for datasets that contained more than one type. This was done to investigate if there was some pattern among the datasets for different contrasts. There was no pattern observed for the Male/Female or MS/Control groups, however some cell types seemed to have a smaller DNA methylation age than others. In Figure 4, it is observed for two different cohorts, that CD4 cells (green) seem to have lower DNA methylation age/age acceleration than the CD14 cells (blue). This is interesting, given that the same individuals were the source of the CD4 or CD14 cells, and therefore the samples are of the same chronological age (paired).

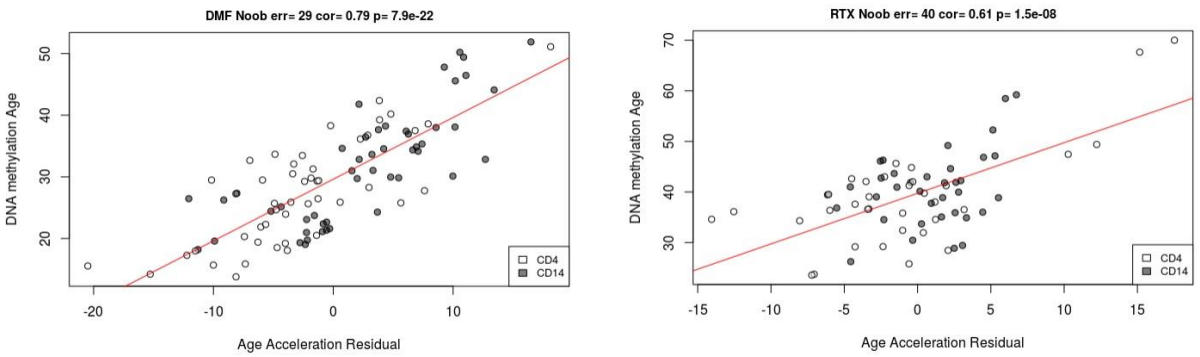


Figure 4. Scatterplots of DNA methylation age versus Age acceleration residual for DMF (left) and RTX (right). CD4 cells are shown in white and CD14 cells are shown in black colour.

A similar pattern is observed in another dataset containing CD8 and CD19 cells. Even though the samples in this case are not perfectly paired, CD8 cells appear to have higher biological age (DNA methylation age/age acceleration residual) than CD19 cells (Figure 5). Noob and Funnorm normalized data scatterplots are not shown due to their high similarity to the raw data plots (showing the same pattern). These patterns are indicative of a clustering between data, however this is not the best way to explore this possibility. The significance of this differences between cell types will be further investigated in later analysis.

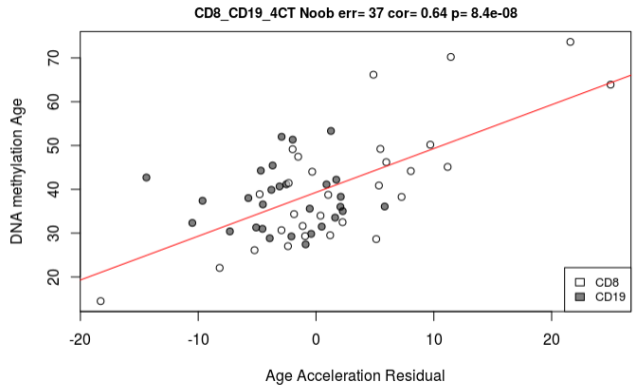


Figure 5. Scatterplot of DNA methylation age versus Age acceleration residual for CD8_CD19_4CT dataset. CD8 cells are shown in white and CD19 cells are shown in black colour.

- ## Contrast level comparison: Bar plots and linear models

The simplest variables with high interest in this study (available for all datasets) are the disease status (either MS or control) and gender (male or female). For this reason, those two variables were chosen to analyse for first. The Broad dataset was selected, due to its highest sample count, which provides more statistical power over the other datasets. In addition, it contains all three measures of age acceleration; age acceleration residual, and the adjusted IEAA and EEAA. Figure 6 is a compilation of bar plots, showing the results for raw data and Noob normalisation.

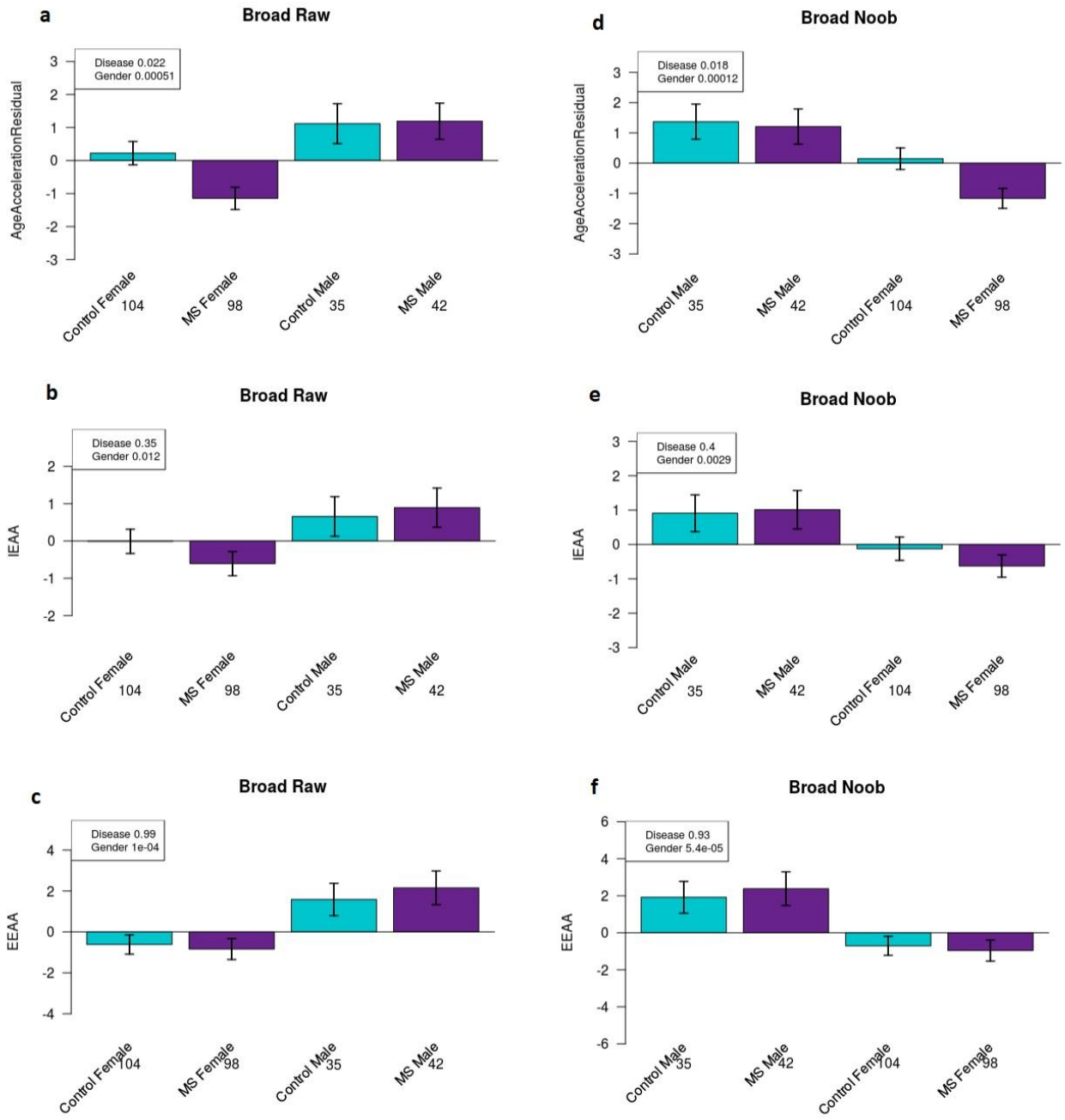


Figure 6. Bar plots for the Broad dataset showing the mean difference between Disease and Gender variables for the three age acceleration measures (residual, IEAA and EEAA). Number of samples is found under each group name. P-values are shown in the legend. a-c) Raw data and d-f) Noob normalized data.

It is observed in Figure 6 that disease status (reference level: control) and gender (reference level: male) are both significantly different contrasts for age acceleration residual (Figure 6a and 6d). However, it is not the same for the adjusted measures IEAA (Figure 6b and 6e) and EEAA (Figure 6c and 6f); in this case only gender is significant. It is shown that samples coming from females appear to have less age acceleration than males, even negative age acceleration. This has been observed before in a study published in 2016, where researchers state that when looking at whole blood samples, the female immune system seems to be younger than the male one (Horvath *et al.*, 2016). In addition, when only looking at females in the case of MS, the ones with the disease seem to have

even lower age acceleration residual (Figure 6a and 6d) that the controls. These patterns were later investigated across other datasets.

The Selected dataset also contains results on IEAA and EEAA, however, this dataset only has female MS patients, so it will only be analysed for other variables. Additionally, there is only age acceleration residual available for the rest of the datasets, since they do not come from whole blood samples (no IEAA and EEAA measures available).

Linear models to answer biological questions

Further analysis was concluded using the epigenetic clock output that was produced using only Noob normalised data in order to answer specific biological questions based on the observations of the preliminary results.

In the preliminary results, it was observed that:

- Females have lower age acceleration than Males.
- Females with Multiple Sclerosis (MS) have lower age acceleration than Control Females.
- Some cell types appear to have lower age acceleration than other cell types.

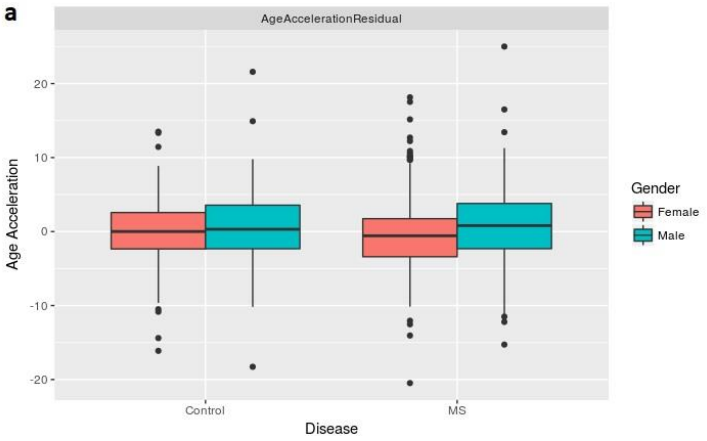
These hypotheses were analysed further and are presented below.

• How does Disease and Gender influence AAR, IEAA and EEAA?

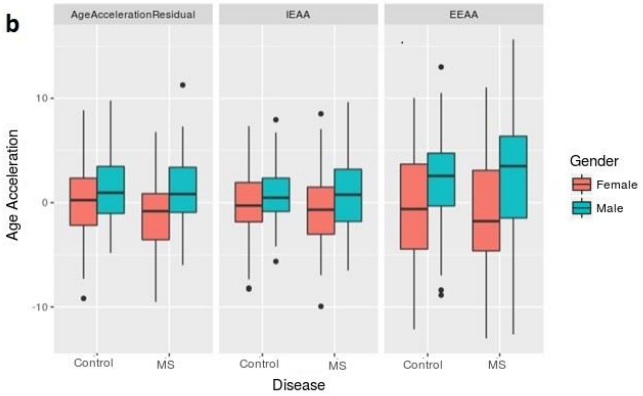
Using "Female" for Gender (due to more samples) and "Control" for Disease as reference levels, the data was analysed for the effect of those two variables to the AAR in all the datasets merged, as well as in the individual datasets (information on the datasets is found in the "Materials and Methods", Table 2), in agreement with the literature (Horvath *et al.*, 2016; Quach *et al.*, 2017).

Not all datasets had the same contrast levels, and therefore, the model had to be adjusted in some cases. In particular, the RTX dataset only has MS patients and in the Brain datasets, there are cases when a subgroup of Gender-Disease status is represented by only one sample (e.g. Brain_ch1: Female-Control). Moreover, in some case the samples were paired, e.g. the different cell types were purified from the same donor (CD14 and CD4 in DMF and RTX). Here, as an initial approach, the role of the individual was not investigated, since it is not possible to have two covariates that hold the same information, in this case Gender and Individual (the variable individual holds information on the individual's gender). In some datasets where more than one cell type was present, the linear model was applied for the variable cell type as well.

In addition, IEAA and EEAA were analysed for the same variables in the merged Broad-Selected datasets. The datasets were merged due to the Selected dataset having only Females with the Disease (MS), and being whole blood samples, as the Broad dataset.



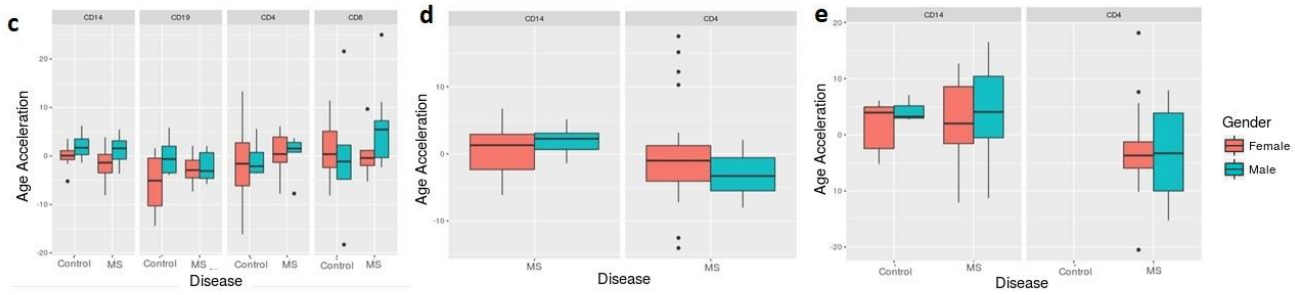
Data	merged all	681
Model	AAR~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.32	0.43
Gender Male	1.05	**0.01**



Data	Broad	279
Model	AAR~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.99629	**0.01834**
Gender Male	1.83212	**0.00012**

Data	Broad	279
Model	IEAA~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.3381	0.40189
Gender Male	1.3541	**0.00289**

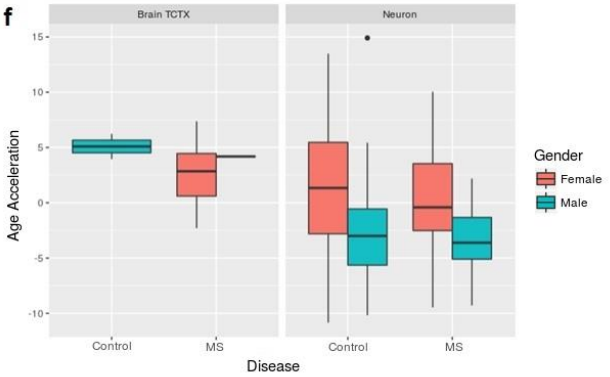
Data	Broad	279
Model	EEAA~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.05586	0.932
Gender Male	3.00172	**5.38E-05**



Data	4CT	137
Model	AAR~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	0.6866	0.4844
Gender Male	2.0288	**0.0423**

Data	RTX	70
Model	AAR~Gender	
Factor	Coefficient	P-value
Gender Male	-0.6993	0.634

Data	DMF	93
Model	**AAR~Disease+Gender**	
Factor	Coefficient	P-value
Disease MS	-2.782	0.235
Gender Male	0.9496	0.564



Data	Brain ch1+2	47
Model	**AAR~Disease+Gender**	
Factor	Coefficient	P-value
Disease MS	-0.9022	0.668
Gender Male	-2.6369	0.21

Figure 7. Box plots showing the distribution of the samples in all datasets (a), in Broad (b), four purified cell types (CD14, CD4_4CT and CD8_CD19_4CT merged) (c), RTX (d), DMF (e), and brain (Brain_ch1 and Brain_ch2 merged) (f). The datasets are grouped by factors Disease (MS, Control) and Gender (Male – teal, Female – red), and Cell or Tissue Type where applicable. The linear model details are shown to the right or under each plot in a table. The tables show the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

When looking at the merged data (all datasets), the AAR seems to be significantly higher for male individuals compared to the females (Figure 7a). This is in accordance with previous work from Horvath et al. (Horvath *et al.*, 2016). However, the disease (MS) does not seem to be significant in AAR according to this data. When analysed in other (merged or not) datasets (Figure 2b-f), the association of Gender to AAR seems to be significant in several datasets and the association with MS was significant only in blood (Figure 7b) but not in purified cell types (Figure 7 c-f). Specifically, the data suggests that females with MS could have lower age acceleration in blood. Possible explanatory factors and confounder may include the different relative fraction of blood cell types in MS as compared to controls and the inter-individual variability.

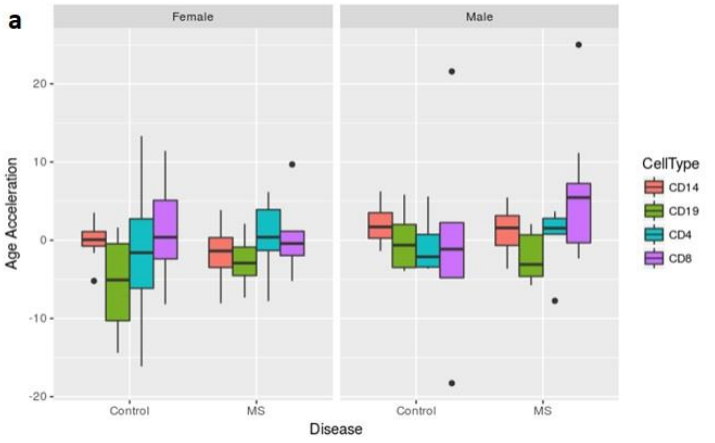
Considering the observations in the preliminary results and observing the differences between cell types shown in the graphs of Figure 7c-e, additional models were used to examine the possibility of the driver of differences between MS/Control and Males/Females being behind the differences in cell type AAR.

The brain datasets were not further investigated, due to the intricacy of the individual datasets (small sample size, disproportional gender and disease groups, very small number of bulk tissue compared to sorted neurons), making the statistical analysis unreliable due to loss of power.

- **Cell Type: How does AAR differ among the major blood cell types?**

To answer this question, the datasets CD14, CD4_4CT and CD8_CD19_4CT were considered (Figure 8). These datasets have partially paired data (samples of different cell types belonging to the same individual). After merging the three datasets, AAR was explained by Gender, Disease and Cell type. The random effect of the Individual was considered but was is not shown, since the paired data percentage was very low in this dataset.

It was shown that both the Cell Type and Gender are significant for the differences observed in Figure 3. Particularly, CD19 cells seem to have the lowest epigenetic age acceleration compared to the other three cell types, and also female individuals seem to have lower age acceleration in their cells than males. This is in concordance with previous findings (see above – hypothesis based on Broad dataset – whole blood data).



b

Data	4CT	137
Model	AAR~Disease+Gender+CellType	
Factor	Coefficient	P-value
Disease MS	0.80	0.40
Gender Male	1.99	**0.04**
CD19	-2.56	0.06
CD4	0.28	0.82
CD8	2.57	0.05

c

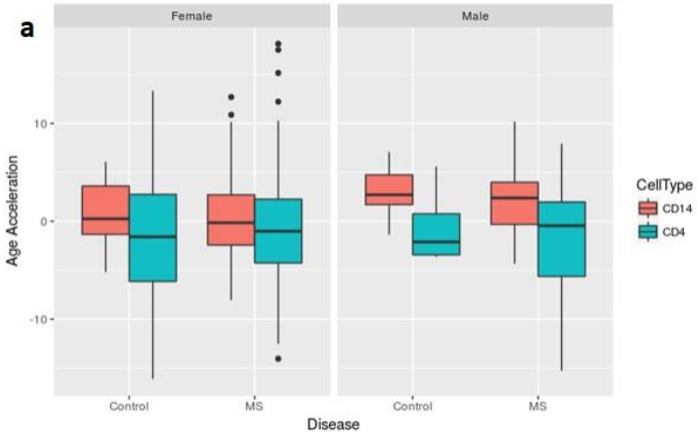
Data	4CT	137
Model	AAR~Disease+Gender+CellType	
Factor	P-value	
Disease		0.40
Gender		**0.04**
CellType		**0.01**

Figure 8. a) Box plot showing the distribution of AAR by the factors: Cell Type (CD14 – light red, CD19 - green, CD4 - teal, CD8 – lilac), Gender (Male, Female) and Disease (MS, Control). b) Linear model for AAR explained by Cell Type, Gender and Disease, showing coefficients for all cell types. c) Linear model for AAR explained by Cell Type, Gender and Disease (after *drop1*). The tables show the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

It is noteworthy that table describing the models in Figure 8c is product of the *drop1* function on a linear model, which removes (drops) each factor of the model separately and provides with various parameters for the statistic performed, among them the p-value (of F test) of that factor. This p-value shows the significance of the previously estimated coefficient of that factor in the linear model. When a factor has only two levels, the p-value is the same as the p-value shown in the linear model summary output, which corresponds to the non-reference level of that factor.

- **Cell Type: How does AAR differ between CD4 and CD14 cells in MS?**

In order to answer this question, datasets containing CD4 and CD14 cells were merged (CD14, CD4_4CT, DMF, RTX). Moreover, for the DMF and RTX datasets, only the baseline samples were selected (before therapy). Since this merged dataset contains samples from the CD14 and CD4_4CT datasets, not all data is paired (only DMF and RTX data is paired). Additionally, it is noteworthy, that there are fewer samples in the control group, since RTX dataset did not contribute any control samples.



b

Data	CD14+CD4_4CT+RTX+DMF	192
Model	AAR~Disease+Gender+CellType	
Factor	Coefficient	P-value
Disease MS	-0.11	0.91
Gender Male	0.62	0.46
CD4	-1.70	**0.03**

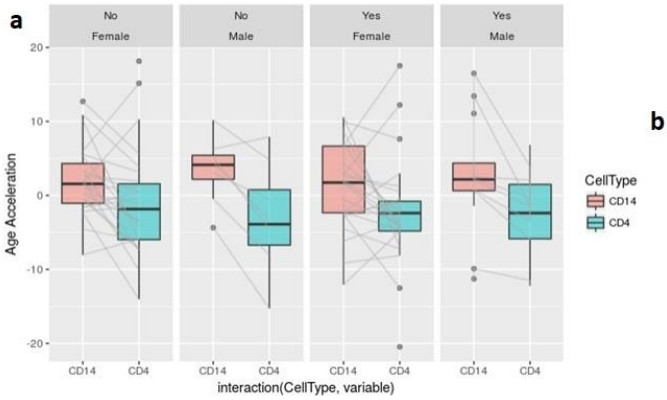
Figure 9. a) Box plot showing the distribution of AAR by the factors: Cell Type (CD14 – light red, CD4 - teal), Gender (Male, Female) and Disease (MS, Control). b) Linear model of AAR explained by Cell Type, Gender and Disease. The table shows the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

When looking at only CD4 versus CD14 cells, it was observed that cell type was statistically significant factor for the differences in AAR among the groups, while Gender and Disease did not appear significant in the linear model (Figure 9b). Particularly, CD4 cells seem to be of lower age acceleration compared to CD14 cells regardless of Gender and Disease status.

As well as in the previous case with the major cell types, the simple linear model was compared with the mixed model, with the Individual as random effect. When comparing two models it was shown that adding Individual as random effect improves the model (data not shown); looking at the Bayesian Information Criterion (BIC) value of the model is reduced with respect to that of the simple linear model and the difference is >10 and significant (p-value of χ^2), which indicates very strong evidence against the simple model. Therefore, the random effect of the individual was more significant for this merged dataset than in the previous case. However, it was decided not to use it in this report, since the data were not completely paired in this case either and adding the individual as random effect did not improve the significance of other factors in the original model.

- ## Therapy: Does it affect AAR in CD4 and CD14 cells?

Seeing the differences of AAR between the CD4 and CD14 cells in a previous model (Figure 9), an additional question was raised; would this difference in AAR be reduced after therapy (treatment with RTX or DMF)? In order to answer this question, the AAR was explained by Cell Type, Gender, and receiving therapy with either medicine (Therapy: yes = after, or no = before) (Figure 10).



b

Data	RTX+DMF	152
Model	AAR~Gender+CellType+Therapy_YN+(1|Individual)	
Factor	Coefficient	P-value
Gender Male	0.09	0.94
CellType CD4	-4.08	**1.78E-07**
Therapy_YN Yes	-0.89	0.22
(1|Individual)		**2.75E-12**

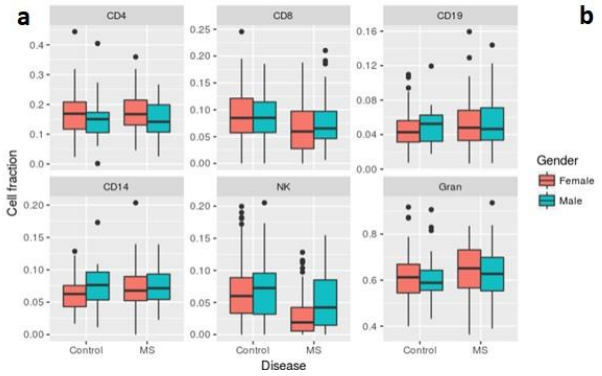
Figure 10. a) Box plot showing the distribution of AAR by the factors: Cell Type (CD14 – light red, CD4 - teal), Gender (Male, Female) and Therapy (Yes = after, No = before). Grey lines indicate the paired data. b) Linear mixed model for AAR explained by Cell Type, Gender and Therapy, incorporating the random effect of the Individual. The table shows the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

Since this data is completely paired (for every individual there is CD4 and CD14 cell samples, before and after therapy, apart from some DMF samples, explained in "Data" of "Materials and Methods"), a paired box plot was considered for the better visualisation of the data. Here, the pattern is obvious; in every group, most of the lines are descending from the CD14 to the CD4 cells, indicating the lower age acceleration of the CD4 cells. This effect is regardless of the Gender or Therapy status, as confirmed by the model as well (Figure 10b). Moreover, the addition of the random effect of the individual in a mixed model, improved the simple model, as expected (lower BIC, >10 units difference, data not shown). Therefore, the linear mixed model was chosen as more appropriate to show in this case.

- **Cell type: How do the individual cell fractions differ based on Gender and Disease?**

So far, the results have been based on the purified cell types available in specific datasets. However, it was interesting to investigate how the fractions of each cell type differ among individuals of different gender or disease status, when whole blood is taken as sample, which comprised of not only the previously investigated cell types, but also natural killer (NK) cells and granulocytes (Gran). The cell type fractions can be estimated using an R package (*FlowSorted.Blood.450K*) (Jaffe, 2016), by DNA methylation markers specific to cell types. Especially since it was previously shown that different cell types can have different age acceleration, it can be assessed if the driving factor in the effect of Gender on AAR is the cell counts (fractions) or another factor that Males and Females differ upon (Figure 11).

a



b

Data	Broad	279
Model	CD4~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	0.01	0.43
Gender Male	-0.02	**0.01**

Data	Broad	279
Model	CD19~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	0.01	**0.05**
Gender Male	0.00	0.35

Data	Broad	279
Model	CD14~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	0.01	0.06
Gender Male	0.01	**0.04**

Data	Broad	279
Model	NK~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.03	**1.80E-09**
Gender Male	0.02	**0.01**

Data	Broad	279
Model	CD8~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	-0.02	**1.18E-03**
Gender Male	0.00	0.70

Data	Broad	279
Model	Gran~Disease+Gender	
Factor	Coefficient	P-value
Disease MS	0.03	**0.02**
Gender Male	-0.01	0.45

Figure 11. a) Box plot showing the distributions for each cell type fraction in the whole blood samples, divided by Gender (Female – red and Male – teal) and Disease (MS, Control). Not that the y axes show percentage and are free scale among the cell types. b) Linear models for each cell type (CD4, CD8, CD19, CD14, NK, Gran) explained by Gender and Disease status. The tables show the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

Based on the box plot of Figure 11, as well as the linear models, it is shown that there are significant differences between the Gender and/or the Disease status for all cell type fractions. In particular, it seems that Females have a statistically significantly higher fraction of CD4 cells, while Males have a higher fraction of CD14 and NK cells. When looking at the Disease status, the Controls seem to have a higher fraction of CD8 cells, while the MS patients seem to have a higher fraction of CD19 cells and Granulocytes, regardless of the Gender.

It is noteworthy, that even though the coefficient values are low, they represent cell fractions changes, which are already measured in percentages (0-1), and all cell fractions but granulocytes are already very low in value to begin with.

- **MS risk factors: Is their contribution to age acceleration significant in whole blood?**

In order to investigate this question, the Broad dataset (whole blood samples) was selected, and known MS risk factors were considered in the model; these were Gender, HLA risk (yes or no, determined as having the risk allele DR15), Vitamin D levels (sufficient or insufficient, with sufficient levels being ≥50 nmol/L), Smoking status (non-smoker, past smoker, and current smoker), and body mass index (BMI) at 20 years of age (>27 being risk) (Hedström, Olsson and Alfredsson, 2012; *Multiple Sclerosis International Federation*, 2016; Wergeland *et al.*, 2016).

a

Data	Broad	279
Model	AAR~Gender+HLA_risk2+BMIat20_27+VitD50+SMOKING	
Factor	Coefficient	P-value
Gender Male	1.83	**2.81E-04**
HLA_risk2 Yes	0.15	0.74
VitD50 Insufficient	0.32	0.54
BMIat20_27 High	0.01	0.99
SMOKING CS	0.44	0.46
SMOKING PS	0.22	0.66
* only Gender significant after drop1		

b

Data	Broad	279
Model	IEAA~Gender+HLA_risk2+BMIat20_27+VitD50+SMOKING	
Factor	Coefficient	P-value
Gender Male	1.33	**0.01**
HLA_risk2 Yes	0.44	0.30
VitD50 Insufficien	0.55	0.27
BMIat20_27 High	0.18	0.85
SMOKING CS	0.62	0.28
SMOKING PS	0.26	0.59
* only Gender significant after drop1		

c

Data	Broad	279
Model	EEAA~Gender+HLA_risk2+BMIat20_27+VitD50+SMOKING	
Factor	Coefficient	P-value
Gender Male	3.08	**6.16E-05**
HLA_risk2 Yes	1.49	**0.03**
VitD50 Insufficient	0.52	0.51
BMIat20_27 High	-0.96	0.54
SMOKING CS	-0.24	0.80
SMOKING PS	-0.01	0.99
* Gender and HLA risk significant after drop1		

Figure 12. Linear models for (a) AAR, (b) IEAA and (c) EEAA, explained by Gender, HLA risk allele, Vitamin D levels, Smoking status and BMI at age of 20. The tables show the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

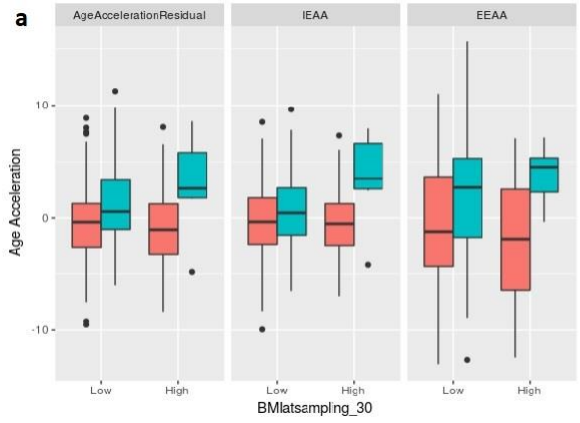
According to the linear models presented in Figure 12, the contribution of risk factors for MS, apart from Gender which has been previously reported throughout this thesis, do not appear to be significant in the three age acceleration measures of the MS patient samples.

After investigating the MS risk factors, it was a logical next step to investigate the aging related factors that might be influencing the epigenetic clock measure in a different way (since the models are based on different DNA methylation markers).

- **Aging factors: How do they contribute to the different measures of age acceleration?**

For this question, the factors considered were: the BMI at the sampling date (low and high, with high being >30), the Gender of the individuals and their Disease status.

a



b

Data	Broad	279
Model	AAR~Disease+Gender+BMIatsampling_30	
Factor	Coefficient	P-value
Disease MS	-0.99	**0.02**
Gender Male	1.83	**1.33E-04**
BMIatsampling_30 High	0.14	0.84

c

Data	Broad	279
Model	IEAA~Disease+Gender+BMIatsampling_30	
Factor	Coefficient	P-value
Disease MS	-0.31	0.45
Gender Male	1.36	**2.93E-03**
BMIatsampling_30 High	0.60	0.35

d

Data	Broad	279
Model	EEAA~Disease+Gender+BMIatsampling_30	
Factor	Coefficient	P-value
Disease MS	-0.10	0.88
Gender Male	3.01	**5.88E-05**
BMIatsampling_30 High	-0.56	0.59

Figure 13. a) Box plot showing the age acceleration for each age acceleration measure in the whole blood samples, divided by Gender (Female – red and Male – teal) and BMI (Low, High). To the right, Linear models for AAR (b), IEAA (c) and EEAA (d), explained by Gender, Disease status and BMI at sampling date. The tables show the data used, the number of samples, the model, the coefficients and p-values for each factor of the model. Significant p-values (<0.05) are marked in bold.

As seen previously in the Broad dataset, the Gender was significant in all age acceleration measures, and the Disease status only in AAR. However, the BMI of the individuals at sampling date did not seem to affect any of the age acceleration measures, even though it is observed in Figure 13 that in the case of high BMI, males seem to have much higher age acceleration (all three measures) than females. However, the samples with BMI at sampling >30, were not as many to provide significance to the overall factor. This factor should be investigated using different parameters perhaps (bigger sample size, more samples for BMI>30). In addition, other variables connected to BMI, such as waist circumference and diet preferences can potentially improve the model. Especially in the case of EEAA, environmental factors play an important role in age acceleration values (Quach *et al.*, 2017) and therefore, more relevant variables need to be used.

Discussion of methods and Workflow proposition

In order to facilitate future analyses, the methods used in this project are discussed below and a workflow is suggested (Table 6).

Table 6. Workflow description and objectives for each step for a comprehensive analysis for the epigenetic clock.

	Step description	*Objectives*
1	Sample and variable collection	Wide selection of phenotypic variables.
2	Experimental process	IDAT files generation.
3	Pre-processing/Normalisation	Noob normalisation.
4	Epigenetic clock	Preparation of files. Submission to the algorithm and output receival.
5	Preliminary analysis	Correlation test, error estimation. Variable distribution observation. Bar plots and t-tests or linear models to investigate differences among groups. Clustering to investigate the existence of discrete groups f observations. Formation of initial hypotheses.
6	Biological questions exploration	Use of the right dataset to test a hypothesis. Visualisation with box plots. Investigation of hypothesis using linear (simple or mixed) models.

Step one is the collection of the samples and variables that describe them; some datasets had a scarce number of variables in the current project (e.g. Brain datasets), making deeper exploration of the dataset difficult. Therefore, more variables need to be obtained in order to correctly form some models and assess differences between groups. Particularly, more “aging variables” (Quach *et al.*, 2017) associated with the epigenetic clock, would help with creating more inclusive linear models that provide a better description of the variability among individuals; e.g. not only BMI at sampling date, but also waist circumference, alcohol consumption, dietary habits etc. In addition, knowing that the IEAA and EEAA are influenced by different factors, having those factors available for more datasets, as well as more datasets with whole blood samples, would provide a better base for comparing those two epigenetic age acceleration measures. Lastly, larger datasets and paired samples would provide more power to the analyses.

Step two is the handling, processing and experimental pipeline followed. Nothing to add to this part.

Step three is the pre-processing and normalisation. According to the results obtained in this project, Noob normalisation was found to perform equally to, or better than Funnorm normalisation and is therefore proposed. However, using raw data for the epigenetic clock analysis is another option,

should someone prefer to not normalise the data. It is noteworthy, that the data should be pre-processed/normalised using only one option among all datasets and future analysis, therefore the initial selection of normalisation method is pivotal, to avoid introducing unwanted variation.

Step four is correctly annotating the files to be submitted for the epigenetic clock analysis and selecting the correct options depending on the tissue type submitted. Every dataset is required to be normalised by the modified BMIQ, incorporated to the epigenetic clock analysis tool. In addition, only whole blood samples should undergo the advanced analysis on the online algorithm; e.g. when submitting the batch Selected_CD14, the whole blood analysis option was selected, and the algorithm returned the additional variables not only for the Selected, whole blood samples, but also for the CD14 purified cells. However, the latter are biased, since the advanced analysis depends on all blood cell fractions and cannot give reliable results for any other tissue or purified cell type.

Step five, after receiving the epigenetic clock algorithm output, it is pivotal to do an observational analysis for the fit of the data. This is done by correlation tests among the variables of interest; DNAmAge and Chronological Age must correlate highly, while the age acceleration residual must not correlate at all with Chronological age, since age has been regressed out. In addition, the error of the model predictions must be estimated, in order to assess the trustworthiness of the results; this is done by calculating the median of the absolute difference between the DNAmAge and the corresponding Chronological age of a sample. Additionally, by creating the scatter plots that correspond to the correlation tests, one can observe the distribution of the data of the output, in order to get an initial insight about the data and the angle on which one can proceed. Lastly, quick observations can be made by using bar plots to visualise and highlight differences and linear models (or t-tests) to confirm those differences in age acceleration, using the variables most used when making associations with differences in the epigenetic clock; gender and disease status (with disease or control samples) (Horvath *et al.*, 2016; Quach *et al.*, 2017). Lastly, clustering can be performed on the data, to investigate if there is clear separation between groups in the data. However, clear separation of data might not be observed in all datasets. After investigating the data through all the aforementioned options, initial hypotheses can be made, so further analysis can be conducted.

Step six comprises the additional analysis based on more concrete questions targeting biologically meaningful answers. To this end, the right dataset needs to be chosen to answer a specific question. Not all datasets contain the information needed to perform the correct statistical analysis. Using box plots to visualise the differences in distributions among groups of variables (contrasts) can prove to be insightful. Box plots offer a better overview of the data, since the distribution can be visualised, compared to the bar plots where only the mean and the standard error are presented. Moreover, using linear models that correspond to specific questions can confirm or negate a certain hypothesis. In addition, the use of linear mixed models can prove useful, although one should interpret the significance of the linear mixed model with care. Additionally, one should consider where adding a random effect makes sense in the data biologically; this is logical, since any factor added to the model could overfit the data, while improving the model statistically.

Methods not used

At this point, it is appropriate to briefly refer to methods that were considered but not used for this project.

- **Normalisation methods**

Due to the different probe chemistry (design – type II and type I probes) and two colour dyes (red/green) used in the 450k and EPIC arrays, various normalisation methods exist in order to get comparable baseline between the two probe types and proceed to investigate differential methylation in the samples (Marabita *et al.*, 2013; Triche *et al.*, 2013; Dedeurwaerder *et al.*, 2014; Morris and Beck, 2015; Wang *et al.*, 2015; Cazaly *et al.*, 2016; Liu and Siegmund, 2016; Wright *et al.*, 2016; Shiah *et al.*, 2017). This is because different methods correct for different aspects of technical variability and probe intensity variation (probe types II and I). Table 7, below, provides information on the still relevant (continuously found to be performing well) and most widely used normalization methods and what they correct for. All these methods are available via R packages.

Table 7. Normalization methods considered but not used for the epigenetic clock analysis. The table includes the method name, the main objectives and some relevant details, the type of normalization (within or between-array), and the type of data normalized (raw intensities of β values) with each method.

Method	*Objectives*	*Details*	*Normalization*	*Data normalized*
Quantile normalization (QN)	Make probe intensity distributions identical among samples.	Better performance in combination with other type of correction (e.g. background).	Between-array	Raw intensities
Stratified Quantile Normalization	QN based on sex chromosomes for male-female samples. Outlier function to remove zeros.	Employed in *minfi* as extra step to normal QN.	Between-array	Raw intensities
Subset Quantile Normalization (SQN)	Make CpG subsets for different CpG class and apply normal QN.	Same biological features will result in same probe variation.	Between-array (also involves within-array)	Raw intensities
Subset-quantile within array normalization (SWAN)	Subset of probes used to create quantile distribution. Subsets created for type II and type I probes. Remaining probes adjusted to subsets.	Probes with equal CpGs will have equal distribution even if they have different design.	Within-array	Raw intensities
Beta-mixture quantile dilation (BMIQ)	Adjust type II to type I probe distribution.	Done by epigenetic clock online tool (modified).	Within-array	β values
Dasen	Adjust background. Between array QN separately on type II and type I probes.	Combination of two methods, Noob and QN.	Between-array	Raw intensities

Even though all methods were developed to correct for technical variation, BMIQ and SWAN are mainly producing within-array (sample) normalization, while the others produce between-array normalization.

Previous work supports that BMIQ (Teschendorff *et al.*, 2013) is one of the most useful methods for normalizing the data (Marabita *et al.*, 2013; Dedeurwaerder *et al.*, 2014; Wang *et al.*, 2015). However, a modified version of this method is applied by the epigenetic clock online tool, and therefore was not used in the pre-processing/normalization step of this project.

As seen in Table 7, QN is a fundamental method of normalization of type II and type I probes. However, it is best applied together with another correction step (Marabita *et al.*, 2013). Stratified QN is not considered useful in the case of the epigenetic clock analysis, since gender of the individuals is involved in another way (probes linked to sex chromosomes are among the 353 CpGs and are used for quality checks in the epigenetic clock – see Step 5, below). SQN is the least favoured method for normalizing intensities between type II and type I probes, since type I probes are adjusted to type II, which are known to have greater variation (Liu and Siegmund, 2016). SWAN corrects for probe design bias, similarly to BMIQ (Liu and Siegmund, 2016) and will not be used in this project, since this type of normalization is implemented in the epigenetic clock tool.

Dasen (Pidsley *et al.*, 2013) is another favourable method that has been shown to perform well, reducing between-sample variability with great efficiency, similarly to FunNorm (Liu and Siegmund, 2016). Dasen corrects for multiple biases present in the 450k and EPIC methylation arrays (background intensity, type II and type I probes) and is therefore recommended in several recent studies (Liu and Siegmund, 2016; Fortin, Triche and Hansen, 2017; Shiah *et al.*, 2017). However, since FunNorm was used in this project, dasen was excluded.

Finally, batch correction methods like ComBat (Johnson, Li and Rabinovic, 2007) and SVA (Leek *et al.*, 2012), typically used in a differential methylation analysis and in combination with normalisation methods, are not considered at the moment, since the different batches of datasets were submitted separately.

- **Epigenetic age predictors**

In addition to the two epigenetic age models used in this project, more predictors have been developed having the same goal. In fact, this subject was the focus of a recent review by the developer of the epigenetic clock used in this thesis, Horvath, and Raj (Horvath and Raj, 2018). In this review it is explained that other epigenetic age predictors are not as accurate as the Horvath epigenetic clock (or Hannum clock for whole blood data). In addition, even if they perform well in a tissue, usually whole blood, it does not mean that they can predict DNAm Age accurately in other tissues. Only Horvath clock has been trained and validated using thousands of samples across a multitude of tissues and cell types. Finally, a new predictor using DNA methylation values has been recently developed and published, which greatly outperforms the previously developed estimators, called DNAm PhenoAge (Levine *et al.*, 2018). This age predictor has been based on phenotypic markers rather that chronological age, and it regresses ten clinical biomarkers of age (e.g. glucose levels, blood pressure etc.) on DNA methylation levels in blood. PhenoAge is an estimator of mortality and morbidity, and it is suggested to only be used on blood, similarly to Hannum clock. This new estimator would have been a great candidate for this project, however it was published late in the course of the project and therefore could not be used.

- **Statistical analysis and visualisation methods**

The two methods for statistical analysis and visualisation that were ultimately not used in this project was the density plots, to visualise the distribution of a variable among the samples, and the clustering (using *Mclust* package in R), to divide the samples of a dataset into groups for further investigation. Instead of using this approach followed by contingency tables to explore and identify

enrichment of specific variable (factor) levels within the cluster groups, linear models were used to answer targeted questions based on preliminary observations and hypotheses.

4. Discussion and Conclusions

In this project, the conclusions concern both bioinformatics as well as biological aspects. From a bioinformatics point of view, it is concluded that the best normalisation option was Noob normalisation among the three options tested. Although there is no evidence against Noob normalisation in the literature, other studies involving the epigenetic clock algorithm of Horvath reported using only the modified BMIQ normalisation method (Horvath, Mah, *et al.*, 2015; Knight *et al.*, 2016) provided by the online tool (Horvath, 2013), or dasen normalisation (Horvath *et al.*, 2016). Thus, in this project another normalisation method that performs well among various datasets was identified.

For the analysis of the output of the epigenetic clock, R provides several statistical and visualisation options, although it is suggested to use each tool with care and being mindful of the pitfalls it might entail and the interpretation of the results it provides. In various studies using the epigenetic clock algorithm to investigate age acceleration in diseases, visualisation of the results was provided by a combination of scatter plots to show the fit of the predictor model and bar plots to highlight the differences between groups (Horvath *et al.*, 2014, 2016; Horvath and Levine, 2015; Horvath and Ritz, 2015; Horvath, Garagnani, *et al.*, 2015). However, in this project it was considered to add box plots to visualise the distribution of the age acceleration measures among different groups, and additionally to match paired data. Being able to visualise the whole distribution of a variable and not just the mean of a group, can give insight as to the existence of outliers and skewed data.

In the context of age acceleration, several hypotheses were initially formed and investigated further. It was shown in various datasets, as well as the merged data containing multiple tissues, that there is an association between the gender and AAR; initially it was shown in the broad dataset that the female individuals have lower age acceleration (AAR, IEAA and EEAA) than the males. This has been shown before by Horvath et al. (Horvath *et al.*, 2016) and it was attributed to the lower epigenetic age of the immune system of females. This result was confirmed in purified CD14 cells as well.

However, age acceleration has not been previously investigated in the context of MS. In this study, it was shown in the same dataset that female MS patients had an even lower age acceleration than the controls; this was not confirmed in another dataset. However, the Broad dataset has the highest number of samples/Individuals (279/279) and therefore provides more power to statistical analyses which translates to ability to detect more subtle differences between groups. Additionally, the Broad dataset was consisting of whole blood samples, while most of the other datasets available in this project were comprising purified blood cell types, and therefore contained different epigenetic information. Furthermore, it was observed that different cell types belonging to the same batch (dataset) or the same individual had different age acceleration; particularly, when looking at CD4, CD14, CD8 and CD19 cells, CD19 (B cells) seemed to have the lowest age acceleration compared to the other cell types, and gender was also significant in the differences observed in AAR; females had lower age acceleration than males, which was in concordance with the whole blood data. Moreover, looking at datasets with higher degree of paired data (purified cell type samples of the same donor), the CD4 cells had lower age acceleration than the CD14 cells, regardless of Disease status, Gender

and Therapy status. These differences led to the hypothesis that different cell proportions in females and males, could be driving the differences observed in these gender groups. Since it was possible to estimate the cell fractions in the Broad-Selected datasets, Gender and Disease status effect was investigated on these cell fractions. Finally, it was shown that females have a higher proportion of CD4 cells, while males have a higher proportion of CD14. In addition, MS patients seemed to have a higher proportion of CD19 cells compared to the controls, regardless of gender. Given the previous findings about the age acceleration differences between those cell types, this is an indication that the differences observed between females and males, and female MS patients and female controls in whole blood samples might be due to the cell type influence.

These results are significant, since there is no other literature to date that reports the same pattern. Even though previous literature suggests that age acceleration differs between cell proportions (Horvath and Levine, 2015), the authors do not state that one cell type has higher/lower age acceleration than the other. In fact, the expectation is for different tissue types of a specific individual to have similar age prediction (*DNA methylation age and the epigenetic clock*, 2013b; Horvath and Raj, 2018), however this is referring to blood tissue as a whole, compared to other body tissues.

Overall, the analysis on the epigenetic clock output for these datasets revealed some patterns in the preliminary results, which were confirmed by testing the hypotheses. The data tells a story and the results are significant. Following the proposed workflow and improvements, more biologically relevant information could arise from further investigation of all three age acceleration measures in MS. Should this analysis be extended to more datasets, it would help understand the significance of epigenetic age acceleration in specific cell types in the disease.

5. Future directions

Further analysis would be required to confirm the findings of the current study. Other datasets can be used, provided they have a compatible design, with matching tissue, cell type and variables. In addition, improvements on the current analysis were mentioned in “Discussion of methods and Workflow proposition” section. Briefly, more variables are needed in order to make better models to explain differences between groups. Larger datasets and paired data can prove invaluable to the investigation of more subtle differences. These suggestions would also prove significant in similar analyses in other diseases.

Lastly, DNAm PhenoAge on the Broad dataset (whole blood samples) could contribute with additional insight on the morbidity and mortality of the MS patients compared to the controls of this study. However, it is pivotal that the control samples are carefully annotated for other diseases, since they might be relevant to the estimator. Since the controls were selected for not having MS or other inflammatory disease of the brain, but not for other diseases, this falls under the experimental design and planning.

6. Ethical aspects

The datasets used in this study has been originally obtained for other studies. Therefore, there was ethical approval received for each study and for the use of these datasets for epidemiological investigations. DNA was extracted from blood samples given by the individuals after informed

consent. The DNA from neurons and brain matter was extracted from brain tissue of deceased subjects, after brain tissue samples were received following autopsy.

In particular, whole blood of Broad and Selected datasets: EIMS (04-252/1-4, Regionala Etikprövningsnämnden i Stockholm, 2004-09-10). Purified CD4/CD8/CD14/CD19 cells of CD14, CD4_4CT and CD8_D19_4CT datasets: STOPMS II (2009/2107-31/2, Regionala Etikprövningsnämnden i Stockholm, 2010-02-16); 2010/879-31-1 (Regionala etikprövningsnämden i Stockholm, 2010-08-18). Purified CD4/CD14 of DMF/RTX treatment studies: STOPMS II (2009/2107-31/2, Regionala Etikprövningsnämnden i Stockholm, 2010-02-16). Neuronal nuclei (Brain datasets): (2012/1417-31/1, Regionala Etikprövningsnämnden, 2012-09-19); 08/MRE09/31+5 (Wales Research Ethics Committee, 2013-06-18).

Acknowledgements

I would like to thank Maja Jagodic for welcoming me to her research group and giving me the opportunity to work on such an exciting project, and Francesco Marabita, for offering me his knowledge and expertise, guidance and continuous support throughout the thesis. In addition, I want to thank Zelmina Lubovac for her support and encouraging comments, and Björn Olsson for his constructive feedback and engaging discussions. Special thank you to my roommates at CMM (KI), research group colleagues, and all the new friends I made during this master's in Bioinformatics.

References

Aryee, M. J. *et al.* (2014) 'Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays', *Bioinformatics*, 30(10), pp. 1363–1369. doi: 10.1093/bioinformatics/btu049.

Castelo-Branco, C. and Soveral, I. (2014) 'The immune system and aging: A review', *Gynecological Endocrinology*, 30(1), pp. 16–22. doi: 10.3109/09513590.2013.852531.

Cazaly, E. *et al.* (2016) 'Comparison of pre-processing methodologies for Illumina 450k methylation array data in familial analyses', *Clinical Epigenetics*, 8. doi: 10.1186/s13148-016-0241-2.

Conerly, M. and Grady, W. M. (2010) 'Insights into the role of DNA methylation in disease through the use of mouse models', *Disease Models & Mechanisms*, 3(5–6), pp. 290–297. doi: 10.1242/dmm.004812.

Dedeurwaerder, S. *et al.* (2014) 'A comprehensive overview of Infinium HumanMethylation450 data processing', *Briefings in bioinformatics*, 15(6), pp. 929–941. doi: 10.1093/bib/bbt054.

DNA methylation age and the epigenetic clock (2013a).

DNA methylation age and the epigenetic clock (2013b). Available at: https://labs.genetics.ucla.edu/horvath/dnamage/.

European Multiple Sclerosis Platform (2015). Available at: http://www.emsp.org/wp-content/uploads/2015/08/MS-in-EU-access.pdf.

Fortin, J. P. *et al.* (2014) 'Functional normalization of 450k methylation array data improves replication in large cancer studies', *Genome Biology*, 15(11), pp. 1–17. doi: 10.1186/s13059-014-0503-2.

Fortin, J. P., Triche, T. J. and Hansen, K. D. (2017) 'Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi', *Bioinformatics*, 33(4), pp. 558–560. doi: 10.1093/bioinformatics/btw691.

Hannum, G. *et al.* (2013) 'Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates', *Molecular Cell*. doi: 10.1016/j.molcel.2012.10.016.

Hedström, A. K., Olsson, T. and Alfredsson, L. (2012) 'High body mass index before age 20 is associated with increased risk for multiple sclerosis in both men and women', *Multiple Sclerosis Journal*, 18(9), pp. 1334–1336. doi: 10.1177/1352458512436596.

Horvath, S. (2013) 'DNA methylation age of human tissues and cell types', *Genome Biol*, 14(10), p. R115. doi: 10.1186/gb-2013-14-10-r115.

Horvath, S. *et al.* (2014) 'Obesity accelerates epigenetic aging of human liver', *Proceedings of the National Academy of Sciences*, 111(43), pp. 15538–15543. doi: 10.1073/pnas.1412759111.

Horvath, S., Garagnani, P., *et al.* (2015) 'Accelerated epigenetic aging in Down syndrome', *Aging Cell*, 14(3), pp. 491–495. doi: 10.1111/acel.12325.

Horvath, S., Mah, V., *et al.* (2015) 'The cerebellum ages slowly according to the epigenetic clock', *Aging*, 7(5), pp. 294–306. doi: 10.18632/aging.100742.

Horvath, S. *et al.* (2016) 'An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease', *Genome Biology*. Genome Biology, 17(1), pp. 0–22. doi: 10.1186/s13059-016-1030-0.

Horvath, S. and Levine, A. J. (2015) 'HIV-1 infection accelerates age according to the epigenetic clock', *Journal of Infectious Diseases*, 212(10), pp. 1563–1573. doi: 10.1093/infdis/jiv277.

Horvath, S. and Raj, K. (2018) 'DNA methylation-based biomarkers and the epigenetic clock theory of ageing', *Nature Reviews Genetics*. Springer US, pp. 1–14. doi: 10.1038/s41576-018-0004-3.

Horvath, S. and Ritz, B. R. (2015) 'Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients', *Aging*, 7(12), pp. 1130–1142. doi: 10.18632/aging.100859.

Jaffe, A. E. (2016) 'Package "FlowSorted.Blood.450k"', pp. 1–5. doi: 10.1186/147121051386.

Johnson, W. E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118–127. doi: 10.1093/biostatistics/kxj037.

Knight, A. K. *et al.* (2016) 'An epigenetic clock for gestational age at birth based on blood methylation data', *Genome Biology*. Genome Biology, 17(1), pp. 1–11. doi: 10.1186/s13059-016-1068-z.

Kurdyukov, S. and Bullock, M. (2016) 'DNA Methylation Analysis: Choosing the Right Method', *Biology*, 5(1), p. 3. doi: 10.3390/biology5010003.

Leek, J. T. *et al.* (2012) 'The SVA package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics*, 28(6), pp. 882–883. doi: 10.1093/bioinformatics/bts034.

Levine, M. E. *et al.* (2018) 'An epigenetic biomarker of aging for lifespan and healthspan', *bioRxiv*, 10(4), p. 276162. doi: 10.1101/276162.

Liu, J. and Siegmund, K. D. (2016) 'An evaluation of processing methods for HumanMethylation450 BeadChip data', *BMC Genomics*. BMC Genomics, 17(1), pp. 1–11. doi: 10.1186/s12864-016-2819-7.

Lu, A. T. *et al.* (2017) 'GWAS of epigenetic ageing rates in blood reveals a critical role for TERT', *bioRxiv*. doi: 10.1101/157776.

Marabita, F. *et al.* (2013) 'An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform', 2294(December 2015), pp. 333–346. doi: 10.4161/epi.24008.

Marioni, R. E. *et al.* (2015) 'DNA methylation age of blood predicts all-cause mortality in later life', *Genome Biology*, 16(1), pp. 1–12. doi: 10.1186/s13059-015-0584-6.

MethylationEPIC BeadChip by Illumina (2017). Available at: https://www.illumina.com/techniques/microarrays/methylation-arrays.html.

Morris, T. J. and Beck, S. (2015) 'Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data', *Methods*. Elsevier Inc., 72(C), pp. 3–8. doi: 10.1016/j.ymeth.2014.08.011.

Multiple Sclerosis International Federation (2016). Available at: https://www.msif.org.

National MS Society (no date). Available at: https://www.nationalmssociety.org/.

Perna, L. *et al.* (2016) 'Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort', *Clinical Epigenetics*. Clinical Epigenetics, 8(1), pp. 1–7. doi: 10.1186/s13148-016-0228-z.

Pidsley, R. *et al.* (2013) 'A data-driven approach to preprocessing Illumina 450 K methylation array data', *BMC Genomics*, 14, p. 293. doi: 10.1186/1471-2164-14-293.

Quach, A. *et al.* (2017) 'Epigenetic clock analysis of diet, exercise, education, and lifestyle factors', *Aging*, 9(2), pp. 419–446. doi: 10.18632/aging.101168.

Rakyan, V. K. *et al.* (2004) 'DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project', *PLoS Biology*, 2(12), pp. 2170–2182. doi: 10.1371/journal.pbio.0020405.

Richardson, B. (2003) 'DNA methylation and autoimmune disease', *Clinical Immunology*, 109(1), pp. 72–79. doi: 10.1016/S1521-6616(03)00206-7.

Sawcer, S. *et al.* (2011) 'Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis', *Nature*, 476, pp. 214–219. doi: 10.1038/nature10251.

Shiah, Y. J. *et al.* (2017) 'Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array', *Bioinformatics*, 33(20), pp. 3151–3157. doi: 10.1093/bioinformatics/btx372.

Stölzel, F. *et al.* (2017) 'Dynamics of epigenetic age following hematopoietic stem cell transplantation', *Haematologica*, 102(8), pp. e321–e323. doi: 10.3324/haematol.2016.160481.

Teschendorff, A. E. *et al.* (2013) 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data', *Bioinformatics*, 29(2), pp. 189–196. doi: 10.1093/bioinformatics/bts680.

Triche, T. J. *et al.* (2013) 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays', *Nucleic Acids Research*, 41(7), pp. 1–11. doi: 10.1093/nar/gkt090.

Wang, T. *et al.* (2015) 'A systematic study of normalization methods for Infinium 450 K methylation data using whole-genome bisulfite sequencing data.', *Epigenetics : official journal of the DNA*

Methylation Society, 2294(June), pp. 37–41. doi: 10.1080/15592294.2015.1057384.

Wergeland, S. *et al.* (2016) 'Vitamin D, HLA-DRB1 and Epstein-Barr virus antibody levels in a prospective cohort of multiple sclerosis patients', *European Journal of Neurology*, 23, pp. 1064–1070. doi: 10.1111/ene.12986.

Wright, M. L. *et al.* (2016) 'Establishing an analytic pipeline for genome-wide DNA methylation', *Clinical Epigenetics*. Clinical Epigenetics, 8(1), pp. 1–10. doi: 10.1186/s13148-016-0212-7.