2021

Variable Selection In Big Data With Applications To Develop A New Epigenetic Clock

Zhenghong Li

VARIABLE SELECTION IN BIG DATA WITH APPLICATIONS TO DEVELOP A NEW EPIGENETIC CLOCK

by

Zhenghong Li

A Dissertation

Submitted in Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

Concentration: Applied Statistics

The University of Memphis

November 2021

DEDICATION

I dedicate my dissertation work to my family. A special feeling of gratitude to my parents who raised me up and gave me a lot of encouragement in my Life. My wife, Wei, who has been a constant source of support during the challenges of graduate school and life. And my lovely kids, Elena and Justin, who are very special to me. This dissertation is also dedicated to the memory of Dr. Seok P Wong who brought me to the Statistics area and was my Master program advisor. I really appreciate his warm suggestions and help in my life.

ACKNOWLEDGMENTS

I would like to thank the following people, without whom I would not have been able to complete this research! My Ph.D advisor Dr. Lih Y Deng, whose insight and knowledge in the subject matter steered me through this research. Dr. Zhaoming Wang who is the principal investigator of this study. I appreciate his kind support for providing me the opportunity to study on this data and all his valuable suggestions for proceeding this project. And special thanks to Dr. Ching-Chi Yang, who meet with me every week and gave thoughtful suggestions of my work which helped my studies to go the extra mile. And many thanks to Dr. Dale Bowman and Dr. Majid Noroozi to be the committee members of my Ph.D study. At last, I would thank to St. Jude Children's Research Hospital where I have worked for thirteen years for its great academic research environment.

Contents

List of Tables

List of Figures

Abstract

Big data is known for 5V's: (1) "volume" with huge quantity/amount (large n) and/or large number of variables (large p), (2) "variety" with various type, nature, and format, (3) "velocity" with ultra-high speed of data generation/collection, (4) "veracity" for its trustworthiness and quality of big data, and (5) "value" for its insights, usefulness and impact. Current computational resources, traditional methodologies and techniques are hard to keep up with the extraordinary volume of data being generated. Therefore, it is challenging to extract useful information from the big data with current computational resources. In this dissertation, we propose procedures to address some of the issues raised with several strategies for some modern variable selection procedures. In particular, we are evaluating various procedures: (1) random sub-sampling so that the sub-data will be "similar" to the original big data, (2) random row partitions so that "all data" will be included, (3) random column partitions to reduce the dimension size for "feasible" model building and/or variable selection while "all columns" can be included, and (4) random matrix partitions is a natural extension using both "row partition" and "column partition". Results from each proposed procedure can be combined via some ensemble methods.

In aging biomarker study, methylation of cytosine residues of cytosine-phosphate-guanine dinucleotides (CpGs) shows strong associations with aging. Several such epigenetic clocks are proposed in the literature. Hannum clock (2013) with 71 CpGs, Horvath clock (2013) with 353 CpGs, Levine clock (2015), and GrimAge clock (2019) with 1,030 CpGs. We will demonstrate that our proposed procedures can be useful in this research area to build a simpler but useful model for ultra-high dimension data. In our study, a total of 2640 SJLIFE participants of European ancestry were included, consisting of 2112 SJLIFE childhood cancer survivors as training data and a separate 528 cancer survivors as validation data. The data includes 689,414 CpGs. This is a clear example of large p (p=689,414) and the sample size n is much smaller. We demonstrate that we can indeed develop a new DNA methylation-based epigenetic clock with much smaller of CpG sites using the proposed procedures.

Chapter 1

Introduction

1.1 Big data and its challenges

Big data is known for 5V's: (1) "volume" with huge quantity/amount (large n) and/or large number of variables (large p), (2) "variety" with various type, nature, and format, (3) "velocity" with ultra-high speed of data generation/collection, (4) "veracity" for its trustworthiness and quality of big data, and (5) "value" for its insights, usefulness and impact.

Current computational resources, traditional methodologies and techniques are hard to keep up with the extraordinary volume of data being generated. Therefore, it is challenging to extract useful information from the big data with current computational resources.

The big data issue always exists in chronological age biomarker studies. Biologically, ageing is the result caused by the accumulation of a wide variety of molecular and cellular damage over time. As age increasing, people become more at risk to disease and death. To more comprehensively understanding the aging, its mechanism has been widely studied in the past. And scientists hope to find a way to slow it down or event reverse the aging process. With this research purpose, it is important to identify a set of aging associated biomarkers.

The telomere is a region of repetitive nucleotide sequences associated with specialized proteins at the ends of chromosome. And it protects the internal regions of the chromosome. The telomere length is negatively associated with aging.[1] It is observed that with the number of cell replication increasing the telomere length is decreasing. However, the high instability in detection and weak correlation with age-related outcomes limits its extension as an unambiguous biomarker of aging.

1.2 Building a new epigenetic clock

Methylation of cytosine residues of cytosine-phosphate-guanine dinucleotides (CpGs) is another DNA-based biomarker associated with age changing.[2] [3] Recently, increasing evidence suggests that dynamic DNA methylation, the most studied durable and reversible epigenetic al-

teration, outperforms other aging biomarkers for its strong correlations with aging and longevity. In the past few decades, there have been a few epigenetic clocks developed based on the CpGs. Hannum developed an epigenetic aging clock including 71 CpGs [4]. Horvath aging clock [5] includes 353 CpGs. In 2015, Levine and his coworkers identified 513 CpGs for a more accurate DNA methylation clock [6]. In 2019, GrimAge clock was developed with 1,030 CpGs [7].

However, there is no study focused on childhood cancer survivors. Cancer survivors receiving surgery, chemotherapy and radiation during their early childhood may have more cellular or molecular damage and cause later physical or mental health problems as well as accelerated aging process. Thus, there is a potential difference of the epigenetic aging clock for the general population and childhood cancer survivors.

The major purpose of our study is to propose and apply different feature selection methods to select the core variables to develop an efficient and accurate DNA methylation-based epigenetic clock epigenetic clock for SJLIFE childhood cancer survivors.

1.3 SJLIFE data and brief outline of the dissertation

A total of 2640 SJLIFE participants of European ancestry were included. These 2640 childhood cancer survivors were stratified randomized into Age/Gender matched training and validation cohort with ratio 0.8 vs 0.2 (sample size: 2112 vs 528). The data includes 689,414 CpGs sites which were processed using Minfi R package with pre-scaled values distribute from 0 to 1. [8]

The detail of our study plan is showed below.

i. Impact of training sample in features selection

Chapter 2 will show the impact of the training data variation in feature selection and how to apply it as well as develop an efficient and accurate DNA methylation-based epigenetic clock for SJLIFE childhood cancer survivors. 6 random split training cohorts (matched by gender and age) are created and the penalization regression method will be used to select the best set of features (CpGs) for each training cohort. There are two popular and well-established penalization regression models: Ridge regression and Lasso regression.

They are both basing on a linear model [9]:

$$Y = \beta X + \epsilon \tag{1.1}$$

However, there is a major difference from the linear model as they apply additional $L2-$ and $L1-$ norm penalty for Ridge and Lasso regression respectively. The beta estimate for a regular linear is showed below [10]:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.2}$$

The issue is that when $n << p$, the $(X^T X)^{-1}$ term does not exist.

In Ridge regression, it introduces a small extra tuning parameter λ and $L2$ penalty into the loss function:

$$Lossfunction = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1.3}$$

where $\lambda \sum_{j=1}^{p} \beta_j^2$ is called $L2 - norm$ penalty

While the beta estimate for Ridge regression can be calculated by introducing the tuning parameter λ which can be presented as follows [11] [12]:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I_{p\times p}^{-1}) X^T Y \tag{1.4}$$

The Ridge regression can only minimize the effect of irrelevant factors but cannot get rid of them, as it cannot shrink the beta coefficients to zero.

Similarly, in the Lasso regression, a tuning parameter λ and $L1$ penalty are included into the loss function:

$$Lossfunction = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{1.5}$$

where $\lambda \sum_{j=1}^{p}|\beta_j|$ is called $L1 - norm$ penalty

However, there is no closed form for the beta estimate for lasso regression, but we can present it below as it minimizes the loss function [13] [12].

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}(\|Y - X\beta\|^2 + \lambda\|\beta\|_1) \tag{1.6}$$

The Lasso regression can shrink the beta coefficients to zeros which is frequently used in variable selection.

In this analysis, a newer technique will be used which combines the Ridge and Lasso regression methods together called the "Elastic Net" regularization regression method. The elastic net regression method is not solely based on $L1$ or $L2$ penalty. It applies the $L1$ and $L2$ penalty together with a weighted parameter α. The Elastic Net method provide us a flexible way to avoid over penalize (lasso) or under penalize (ridge) and find a best balance point between ridge and lasso regression.

The loss function for elastic net regression showed below:

$$Lossfunction = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|) \tag{1.7}$$

Similar to Lasso regression, there is no closed form of the beta estimate for elastic net method. But we can still present it as follows [14] [12]:

$$\hat{\beta}_{elasticnet} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}[\|Y - X\beta\|^2 + \lambda(\frac{(1-\alpha)}{2}\|\beta\|^2 + \alpha\|\beta\|_1)] \tag{1.8}$$

A list of default or user defined alphas will be tested to have smallest "loss" for the model.

The key of the penalized regression is the loss function. The details will be showed in Chapter 2. The elastic net regression will be used to select the CpGs but not build the model. Once the CpGs set is selected, a linear model will be applied to the CpGs to train the prediction model.

ii. Sub-sample Elastic net regression model

To avoid random selection errors, further analysis will be done in Chapter 3. The repeatedly sub-sample elastic net regression using different sampling proportions will be performed by 10 iterations and different selection criteria (different recurrence rate) will be examined to selected core features (CpGs). Besides the adjusted R-squared, the Akaike information criterion (AIC)[15] and Bayesian information criterion (BIC)[16] will be used to evaluate the trained models. The smaller values of AIC/BIC the better the models are. Their formulas are showed below:

$$AIC = 2k - 2ln(\hat{L}) \tag{1.9}$$

$$BIC = kln(n) - 2ln(\hat{L}) \tag{1.10}$$

Where,

$K =$ the number of parameters,

$n =$ the number of data points,

$\hat{L} =$ the maximized value of the likelihood function

AIC rewards goodness of fit but it also includes a penalty that is an increasing function of the number of estimated parameters, which will avoid the over-fitting problem. The BIC is closely related to AIC which is partially based on the likelyhood function and having the penalty term to control for the over-fitting issue. It is believed that the BIC will select the "True model" if the "True model" is included in the candidate models.[17] [18] [19] But with the huge data volumes, it

is unfeasible to find the candidate models including "True model". In real application, it is usually using both AIC and BIC to evaluate the developed models.

iii. Partition Elastic net regression model

Memory issue is always a problem when dealing with ultra-high dimension data. To solve this situation, the partition method will be explored and discussed in Chapter 4. The study cohort will be randomly split into half vs. half partitions by rows and columns as well as by blocks. Then, the elastic net regression will be applied on each partition to select CpGs. After several iterations, a relatively smaller CpGs pool will be created. Further selection by another elastic net regression on the CpGs pool leads to a final CpGs set. Additional comparison will be done to show the partition elastic net regression method is an alternative way to the regular elastic net regression.

iv. Principal component analysis (PCA) and Autoencoder (AE)

The PCA and AE are very popular and widely used in high dimension data reduction. The PCA mathematically transform the large number of variables into equal number of uncorrelated principal components (PCs). The first PC represents the greatest variance of the scaled data project, the second PC represents the second greatest variance and so on. The first few components can be used to either reconstruct the data or train a prediction model. Similar, AE will encode the input variables to user defined less dimension code layer. And the code layer can be used to reconstruct the data or train the model. It is widely believed that if there is a linear association between X and Y, the PCA and AE is similar. In Chapter 5, these two methods will be explored in real data application. Different numbers of PCs (in PCA) and nodes (in AE) will be used to compare the reconstruct errors between these two dimension reduction methods in real CpGs data.

v. Sensitivity analysis using different training cohort

In Chapter 6, an independently split training cohort will be analyzed to build the epigenetic clocks and compare with the initial built clocks to show the robustness of core features selection using our proposed methods. These clocks will be evaluated in validation cohort and health controls as well as comparing their performance with the initial clocks.

Chapter 2

Sub-sampling and penalization regression for variable selection

2.1 Train, Testing and Validation Sample

For a given dataset, it is common to use probabilistic sampling techniques to select a "training sample" from the "population" to build a model and evaluate the model accuracy on the selected "testing sample". For example, we can apply various random sampling strategies (e.g. stratified random sampling) to choose a "good" "training" and "testing sample" to choose a "representative" sample. This is a key element to build a "better" model without the problem of over/under model fitting. In addition, we can also choose another data set for the purpose of "validation" which is commonly used for the purpose of model selection to choose an appropriate model from several competing models. Such sample is called "validation sample".

Since random sampling schemes are used in choosing various samples (training, testing and validation), the model built and its performance are expected to have random variation/fluctuation, especially for data of small/moderate sample sizes. To obtain a more reliable performance measure, we can repeat the whole procedure a few numbers of times and then combine these results by taking a simple average or other advanced methods. With the decreasing computing cost and increasing power of parallel processing, we believe that this should be a standard practice in the future.

2.2 Penalization regression

Penalization regression is widely used in high dimension data analysis to select the variables to build a model. This method will be applied in this study to deal with the ultra-big methylation CpG data.

In peer studies, a strong linear association was detected between CpGs sites and chronical age, which can be presented as a linear model[4] [5] [6] [7]:

$$Chronical\ age = \sum_{i=1}^{n} \beta_i \times CpG_i + Intercept \tag{2.1}$$

To build an epigenetic clock for childhood cancer survivors, the penalization regression method was used for CpG sites selection. By applying a penalty to each variable included in the model, it can lead the coefficients of less contributive variables close to or equal to zero.

2.2.1 Ridge regression

The ridge regression is also called $L2$ regularization which adds "squared magnitude" of coefficient as a penalty term to the loss function. The formula is showed below:

$$Lossfunction = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.2}$$

where $\lambda \sum_{j=1}^{p} \beta_j^2$ is called $L2 - norm$ penalty and the λ is the tuning parameter.

2.2.2 Lasso regression

The lasso regression is also called $L1$ regularization which adds "absolute value of magnitude" of coefficient as a penalty term to the loss function. The formula is showed below:

$$Lossfunction = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.3}$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is called $L1 - norm$ penalty and the λ is tuning parameter.

2.2.3 Elastic net regression

The elastic net regularized regression linearly combines the $L1 - norm$ and $L2 - norm$ penalties. The formula is showed below:

$$Lossfunction = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda[\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|] \tag{2.4}$$

where λ is the tuning parameter and α controls the mix of $L1$ and $L2$ penalty. Using $\alpha = 1$ gives the lasso that we have seen before. Similarly, $\alpha = 0$ gives ridge.

2.3 Using Regular Elastic net regularized regression for SJLIFE CpGs dataset

This section shows how the elastic net regression is applied to the CpGs dataset and its detail analysis.

Elastic net regression analysis details

The 2640 childhood cancer survivors are stratified randomized into Age/Gender matched training and validation cohort with ratio 0.8 vs 0.2 (2,112 vs 528). Table 1 shows the frequency and percentage of the new splatted train and validation cohort.

Table 1. Frequency and percentage of the split training and validation cohort.

	Train cohort N (%)		Validation cohort N (%)	
DNA age	Female	Male	Female	Male
<21	210 (21.17)	222 (19.82)	53 (21.29)	55 (19.71)
21-40	583 (58.77)	671 (59.91)	146 (58.63)	167 (59.86)
≥ 41	199 (20.06)	227 (20.27)	50 (20.08)	57 (20.43)

To consider the impact of different training cohort in variables (CpGs) selection, 6 training/validation cohort (2,112 vs 528) are repeatedly created as discussed above. The elastic net regularized regression will be performed on each of the training dataset to select variables (CpGs).

By fitting the training data into HPC's memory, elastic regression method is applied to the 6 overall training datasets (n=2,112, p=689,414) with 10 folds cross-validation which randomly split the data into 10 parts and take turns to use the 9 parts of the data to predict the 10^{th} part. This procedure is repeated for different α range from 0 to 1 step by 0.1. The CpGs are selected based on lasso regression with repeated selection under different α settings. The best λ gives a minimum mean cross-validated error for each α.

Analysis results of elastic net regression

In the elastic net regression analysis, at $\alpha = 1$ the lasso regression is performed and the penalty is maximized to shrink the beta estimates of the irrelevant features to 0. By testing other α settings, it can avoid over penalized the parameters and confirm the findings. The ridge regression ($\alpha = 0$) does not do the feature selection but can be used to verify the selected parameter by comparing the beta coefficients.

After performing the regular elastic net regularization regression on the 6 independently split training cohort, 602, 589, 604, 595, 611, 598 CpGs are selected respectively for each training data. Figure 1 shows that the CpGs selection variation exists among different training cohorts and there are 141 CpGs repeatedly selected among the 6 CpG sets.

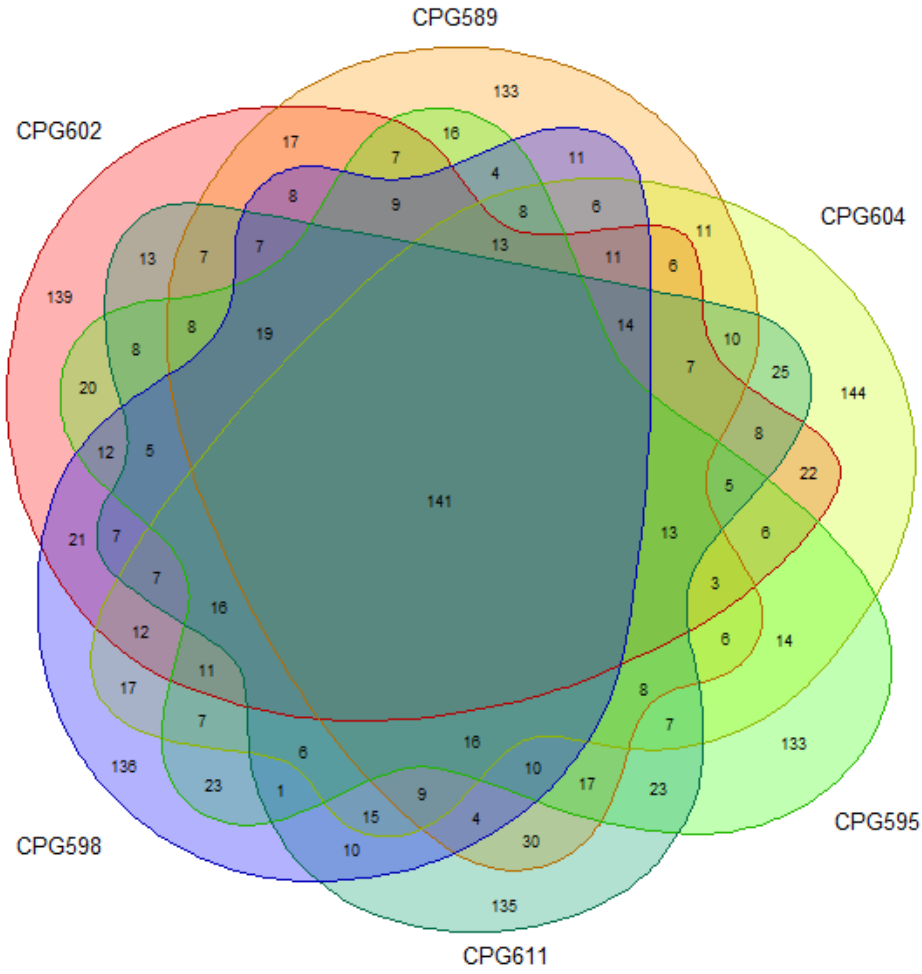

Figure 1. Venn diagram of the 6 selected CpG sets.

Six linear models are trained from the 6 sets of selected CpGs for both the training cohort (n=2,112) and overall cohort (n=2,640). Table 2 below shows their model performance in each cohort.

Table 2. Compare the AIC/BIC among the trained models.

	$Adj - R^2$	AIC	BIC
Training cohort (n=2,112)			
CPG602	0.9928	6334.677	9750.533
CPG611	0.9898	7091.832	10558.59
CPG604	0.9893	7184.779	10611.95
CPG598	0.9891	7223.774	10605.70
CPG595	0.9890	7231.403	10607.67
CPG589	0.9888	7277.357	10619.69
Overall cohort (n=2,640)			
CPG602	0.9913	8838.290	12438.92
CPG611	0.9894	8858.956	12462.50
CPG604	0.9893	8894.721	12457.11
CPG598	0.9891	8936.035	12451.40
CPG595	0.9890	8950.118	12459.60
CPG589	0.9886	9052.512	12526.73

Base on the result, the "CPG602" model performs best while the "CPG589" ranks the bottom compared to other models in both training and overall cohort. The 602 CpGs model gives an adjusted R-squared 0.9928/0.9913 which indicates a perfect model fitness. Thus, the 602 CpGs set is choosen as reference CpGs to build SJLIFE epigenetic clocks.

To evaluate the impact of the selected 602 CpGs on chronical age prediction, the trained 602 CpGs model is compared with another trained model using 513 CpGs included in the phenotypic epigenetic clock – "Levine clock" which calculates the composite "Pheno age" base on the information of age, glucose, C-reactive protein, albumin, creatinine, lymphocyte percentage, mean cell volume, red cell distribution, alkaline phosphatase, and white blood cells. The Levine clock is widely used in epigenetic studies for its good performance in the general population. The comparison is done by applying these two models to the childhood cancer survivors training data.(The result is showed in table 3)

Table 3. Adjusted R^2 for 602 CpGs model and 513 CpGs model.

	$AdjustedR^2$
602 CpGs Model	0.9928
513 CpGs Model	0.9205

The 602 CpGs model has a better adjusted R^2 compared to 513 CpGs model which indicate the 602 CpGs model fits in SJLIFE CPG data better.

Figure 2 shows the scatter plot of the fitted values from 513 CpGs model and 602 CpGs model versus the true chronical age – "DNA sampling age". The 602 CpGs model has a narrower band comparing with the 513 CpGs model which indicate a lower prediction error.

Thus the 602 CpGs model is a more accurate epigenetic clock for SJLIFE survivors.

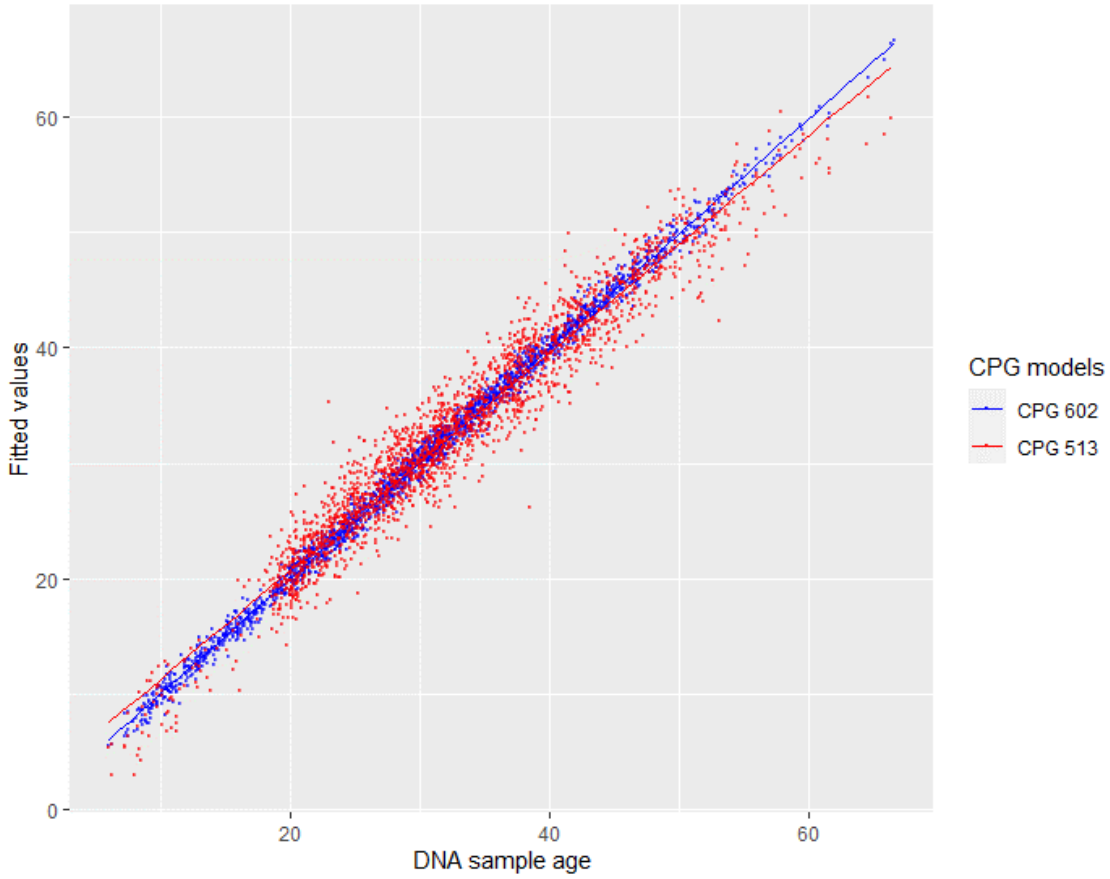

Figure 2. 602 CpGs model vs 513 CpGs model.

Chapter 3

Variables selection procedure with proposed sub-sampling and ensemble method

In big data analysis, a regular elastic net regression may not be able to avoid the redundant variables especially for the case with very large number of parameters but relative small number of observation. In this case, a regular elastic net regression method may over-select the variables and leads to a over-fitting model.

To solve this over-fitting problem and further reduce the number of selected variables, the sub-sampling and train techniques with 10 replications is proposed.

3.1 Foundations for sub-sampling and ensemble method

The analysis in sections 2.3 is based on the 2,112 patients' training data and 689,414 parameters. Though the adjusted R^2 of the trained 602 CpGs model can reach 0.9928, there there are too many variables in the model and may be reduced further.

In this chapter, the proposed sub-sampling and ensemble method will be applied to the CpGs dataset to select the core features as well as build the improved SJLIFE epigenetic aging clock. The original 2,112 SJLIFE patients in training cohort are treated as "Population data", and the 602 CpGs as reference CpG set.

There are a few key foundations in sub-sampling and ensemble analysis:

i. The key important features will be repeatedly selected in various sub-sample of the training cohort.

ii. Large sample size tends to over-select "significant" variables which may not have "practical impact".

iii. The more important features should show up with smaller train sample.

iv. The stricter selection criteria (higher recurrence rate) the more important core features will be selected.

3.1.1 Sub-sampling and ensemble analysis procedure

The sub-sampling and ensemble method has **4 steps**:

i. Random sub-sampling a certain proportion (e.g. 0.8, 0.5, 0.2) of the training cohort.

ii. Perform Elastic net regression on each selected sub-sample.

iii. Repeat step (i) and step (ii) several times (e.g. 10 times) and select the features basing on recurrence rate (e.g. 100%, 90%).

iv. Fit the prediction model based on the selected features and evaluate the models in the test/validation cohort.

It is expected that a small training sample will repeatedly select the core features (CpGs) and the selected CpGs will be the subset of the 602 CPG sites. A cut-off of the CpG recurrence times or selecting CpG outside the 602 CpGs will be considered as a stopping sign.

After the CpGs are selected, the 2112 patients are re-split into the train and test cohort (0.5:0.5) and we use the train data to train a linear model while we use the test data to validate the model.

CpGs selected with 10 recurrence (strictest selection criteria)

Table 4 shows the number of CpGs selected using the strictest selection criteria (CpGs are selected every time among the 10 replications) for different sampling proportion and their model performance.

The sample proportion 0.8 with 10 recurrence of selected 93 CpGs produced the best model. The adjusted R^2 of trained model is 0.9764 and the RMSE (root mean squared error) of the test data is 1.70 which means the predicted chronical age in test cohort may differ ±1.70 years from true age.

The sample proportion 0.5 with 10 recurrence of selected 29 CpGs also leads to a predict model with good of fit. The adjusted R^2 of trained model is 0.9565 and the predicted chronical age may differ ±2.31 years from true age in test cohort.

The sample proportion 0.2 with 10 recurrence of selected 7 CpGs. These 7 CpGs trained a model with adjusted R^2 of 0.89 and the RMSE in test cohort is 3.72. These 7 selected CpGs can

be explained as "core CpGs" as we reduced the data from 689,414 CpGs to only 7 CpGs but we still get an efficient model with acceptable prediction performance.

Table 4. Models comparison between different sample proportion with 10 recurrence of selected CpGs.

Sample proportion	Number of CpGs selected	Trained model Adjusted R^2	AIC	MSE In Test data	RMSE In Test data
0.8	93	0.9764	4349.437	3.671217	1.916042
0.5	29	0.9565	4830.773	5.906927	2.430417
0.2	7	0.8885	5786.153	14.82109	3.849817

Reference: 602 CpGs model with $adj - R^2 = 0.9928$, AIC=6334.677.

In figure 3, it shows the relationship between the 602 CpGs set (Reference set) and the 93, 29, 7 CpGs sets selected from above process. We can see the 93 CpGs is a subset of 602 CpGs (Reference set) and it contains 27 CpGs from the 29 CpGs set. The 7 CpGs set is the "core CpGs" which are contained in all the 602, 93 and 29 CPG sets.

These results support the expectation that the smaller train cohort will select fewer but more important CpGs. And the sub-sampling elastic net regression is an efficient way to reduce the number of features to be selected.

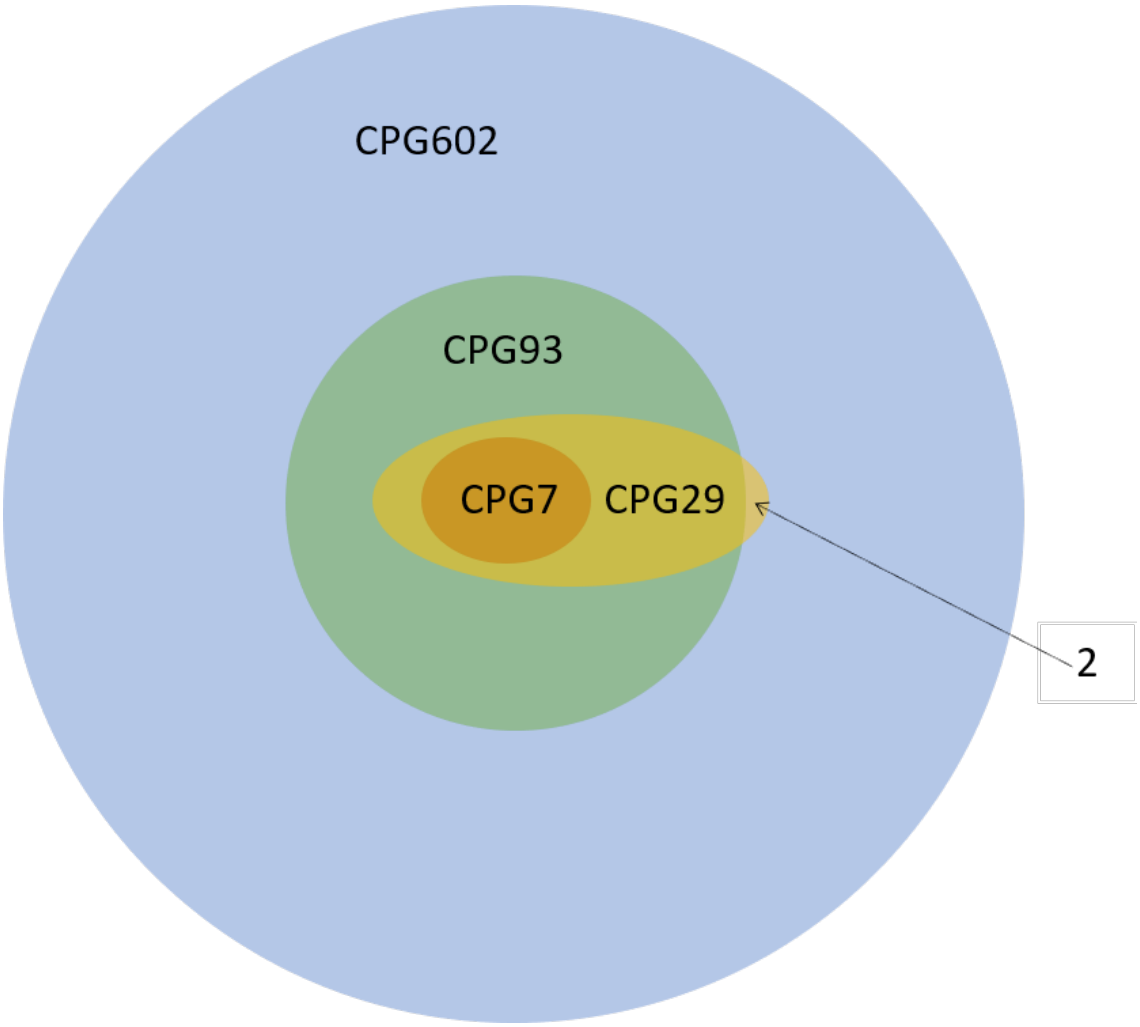

Figure 3. Relationship of the selected CPG sets with 10 recurrence times.

3.1.2 Impaction of recurrence times in variables selection

The recurrence times out of the 10 iterations plays an important role to control the redundant variables. When the cut-off of the recurrence times decrease, more CpGs will be selected. However the opportunity of selecting redundant CpGs will also increase. In the tables and figures below show the change of selected CpGs, trained model and CpGs sets relationship.

i. When select the CpGs with 9 recurrence out of 10 iterations (recurrence rate:90%):

Table 5. Models comparison between different sample proportion with 9 recurrence of selected CpGs.

Sample proportion	Number of CpGs selected	Trained model Adjusted R^2	AIC	MSE In Test data	RMSE In Test data
0.8	151	0.9807	4093.325	3.145465	1.773546
0.5	49	0.9658	4608.031	4.921241	2.218387
0.2	10	0.9106	5583.768	10.93713	3.307133

Reference: 602 CpGs model with $adj - R^2 = 0.9928$, AIC=6334.677.

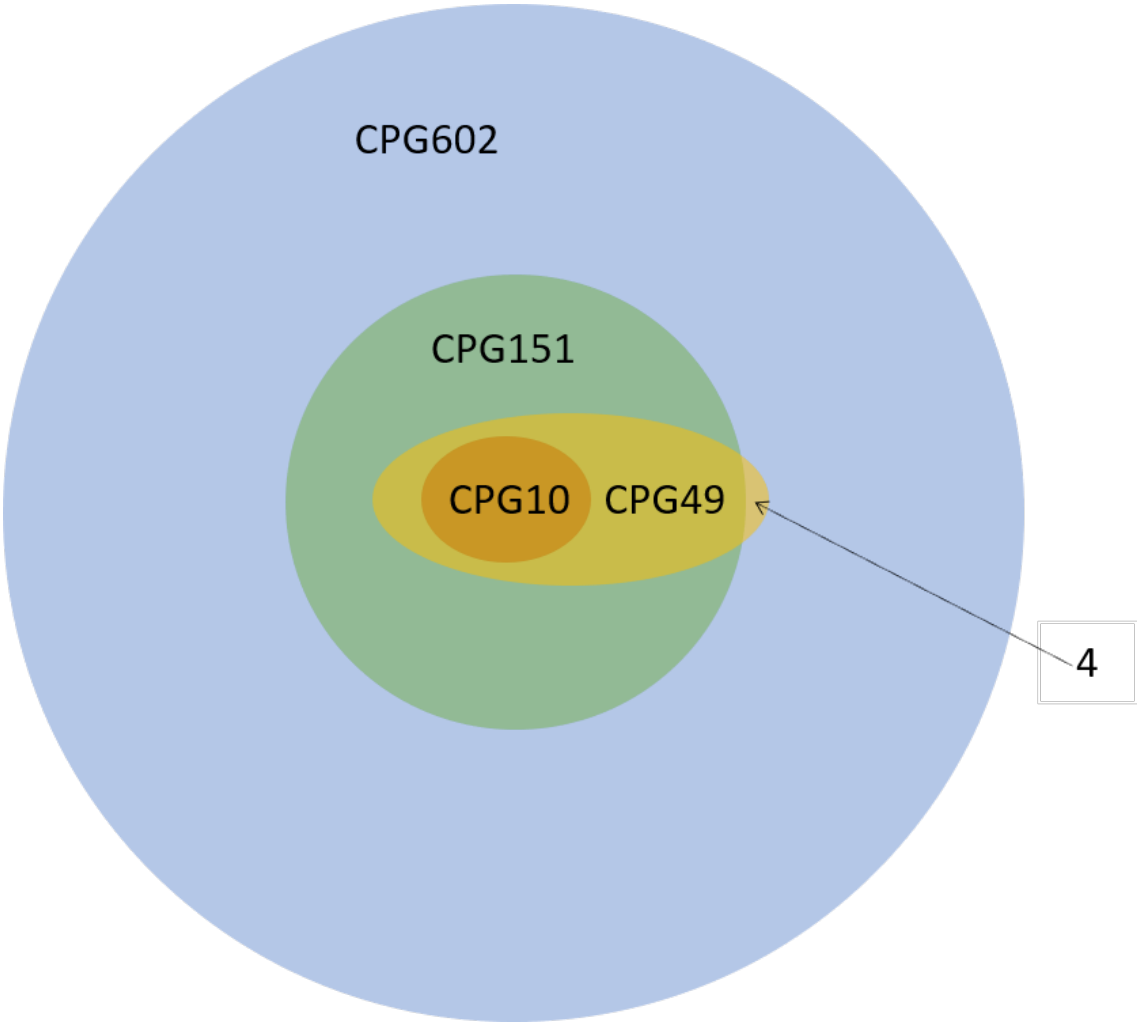

Figure 4. Relationship of the selected CPG sets with 9 recurrence times.

Using 9 recurrence as cut-off to select CpGs, 151 CpGs are selected for sample proportion 0.8, 49 CpGs arere selected for sample proportion 0.5 and 10 CpGs are selected for sample proportion 0.2. And the trained models' performance is comparable to the results using 10 recurrence as cut-off.

The 151 CpGs are subset of the reference CpG set (602 CpGs) and it contains the 49 CpGs set. The 10 CpGs set is the "core CpGs" which are contained in all the 602, 151 and 49 CPG sets.

ii. When select the CpGs with 8 recurrence out of 10 iterations (recurrence rate:80%):

Table 6. Models comparison between different sample proportion with 8 recurrence of selected CpGs.

Sample proportion	Number of CpGs selected	Trained model Adjusted R^2	AIC	MSE In Test data	RMSE In Test data
0.8	206	0.9837	4000.398	2.934534	1.713048
0.5	76	0.9749	4347.531	4.522676	2.126658
0.2	16	0.9467	5084.595	7.842959	2.800528

Reference: 602 CpGs model with $adj - R^2 = 0.9928$, AIC=6334.677.

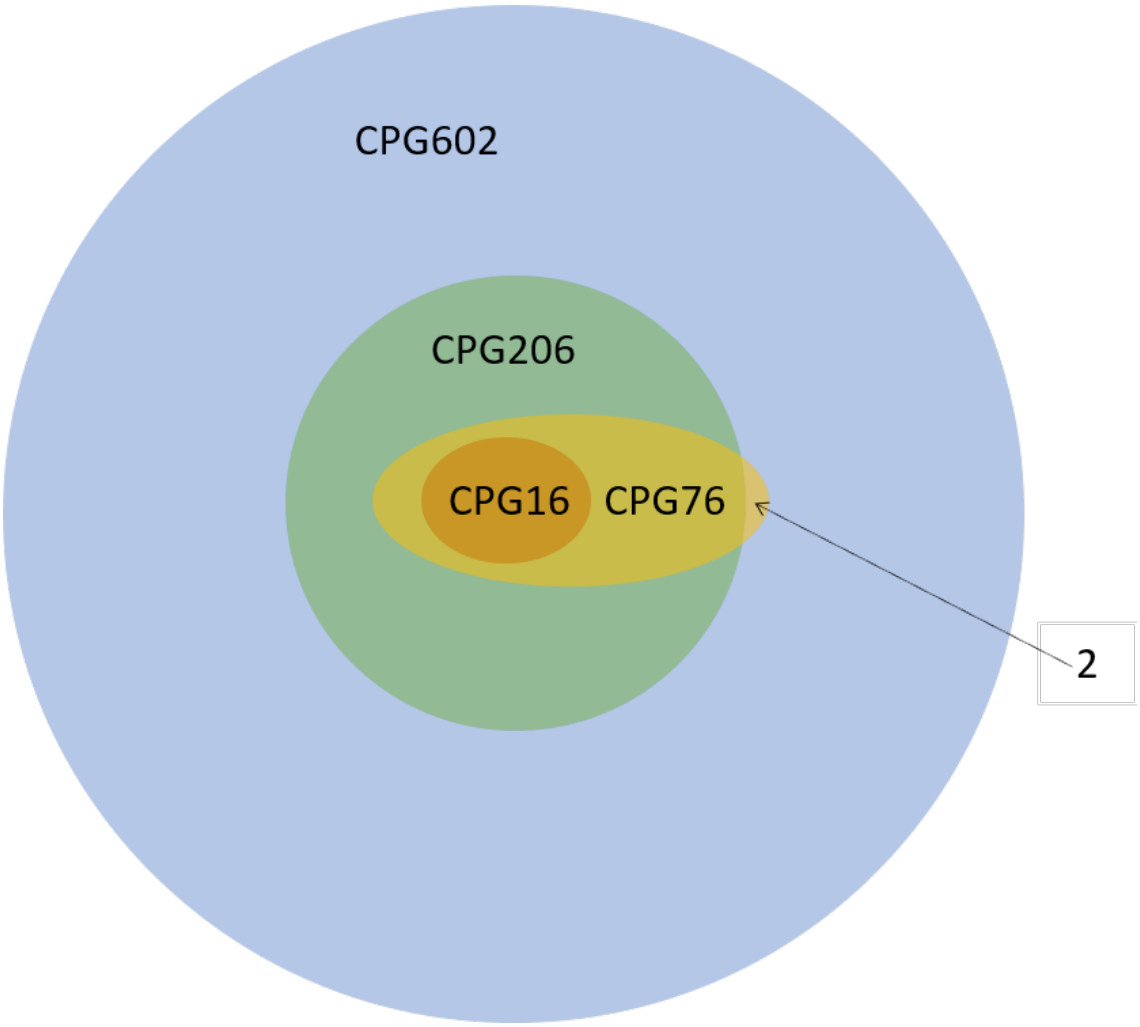

Figure 5. Relationship of the selected CPG sets with 8 recurrence times.

Similar using 8 recurrence times as cut-off 206 CpGs, 76 CpGs and 16 CpGs are selected for sample proportion 0.8, 0.5 and 0.2 respectively. And the trained models have improved model performance as more CpGs are selected. The CpGs sets follow the same relationship as we showed above with the sample proportion decrease the selected CpGs are the subset of the result from the bigger sample proportion.

iii. When select the CpGs with 7 recurrence out of 10 iterations (recurrence rate:70%):

Table 7. Models comparison between different sample proportion with 7 recurrence of selected CpGs.

Sample proportion	Number of CpGs selected	Trained model Adjusted R^2	AIC	MSE In Test data	RMSE In Test data
0.8	256	0.9863	3829.245	2.644145	1.626083
0.5	93	0.9731	4411.879	3.836825	1.958782
0.2	24	0.9522	4954.905	5.927383	2.434622

Reference: 602 CpGs model with $adj - R^2 = 0.9928$, AIC=6334.677.

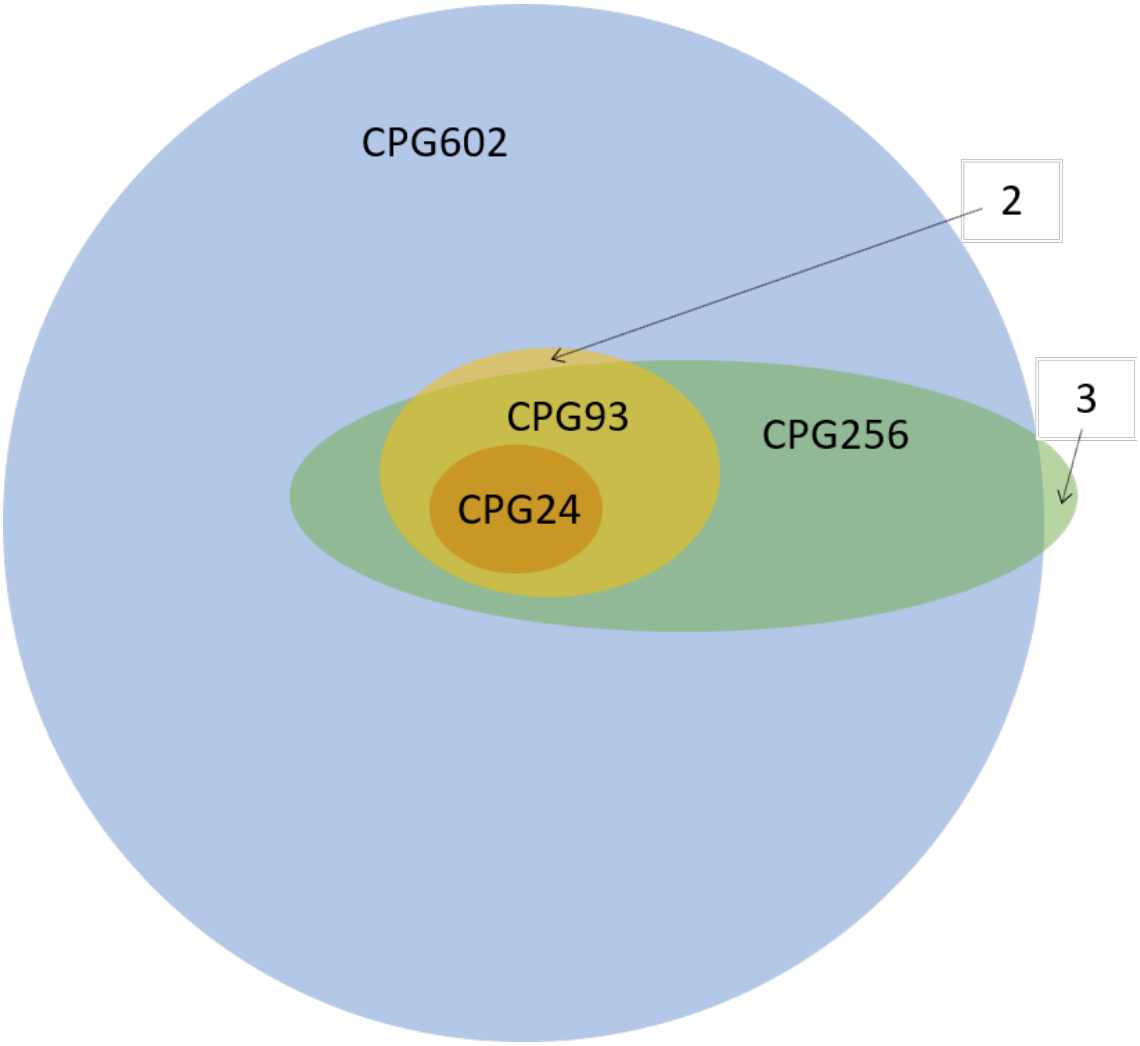

Figure 6. Relationship of the selected CPG sets with 7 recurrence times.

From the figure 6, there are 3 redundant CpGs are selected out of the reference CpG set (602 CpGs). It indicates the 70% recurrence rate may be a potential turning point for controlling the redundant CpGs. To confirm this turning point, we reduce the cut-off recurrence rate to 60% and plotted the Venn-diagram for the selected CpGs sets in figure 7.

When the cut-off recurrence rate decrease to 60% more redundant CpGs (16 CpGs) are selected. This result confirms the 70% recurrence rate is a turning point. It can be concluded that in the sub-sample elastic net regression, to avoid selecting redundant CpGs, CpGs recurrence rate should be at least 80% to keep them in the final sets.

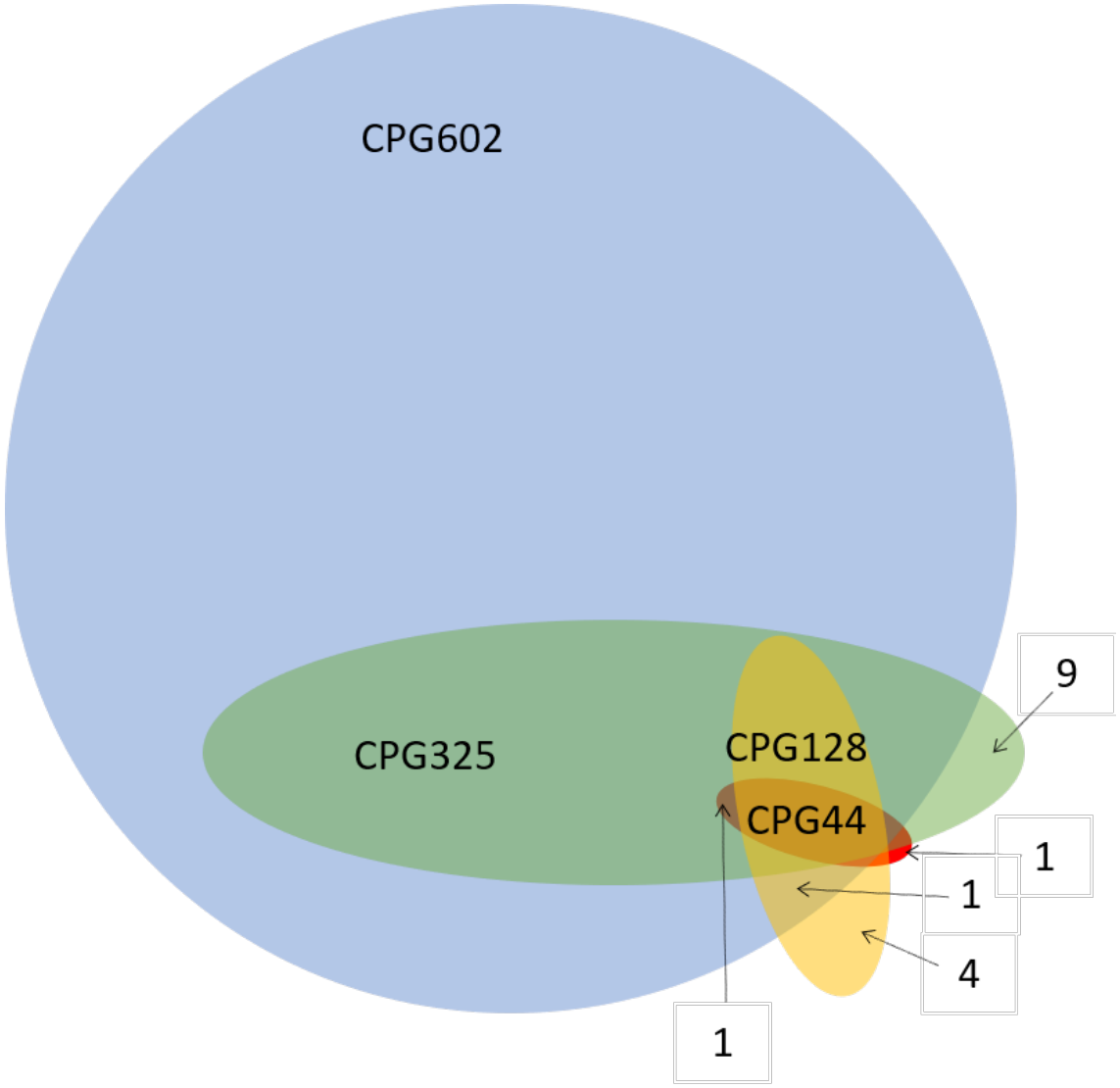

Figure 7. Relationship of the selected CPG sets with 6 recurrence times.

An overall comparison of the adjusted R^2 for the trained models based on different recurrence rates and sample proportions is showed in figure 8.

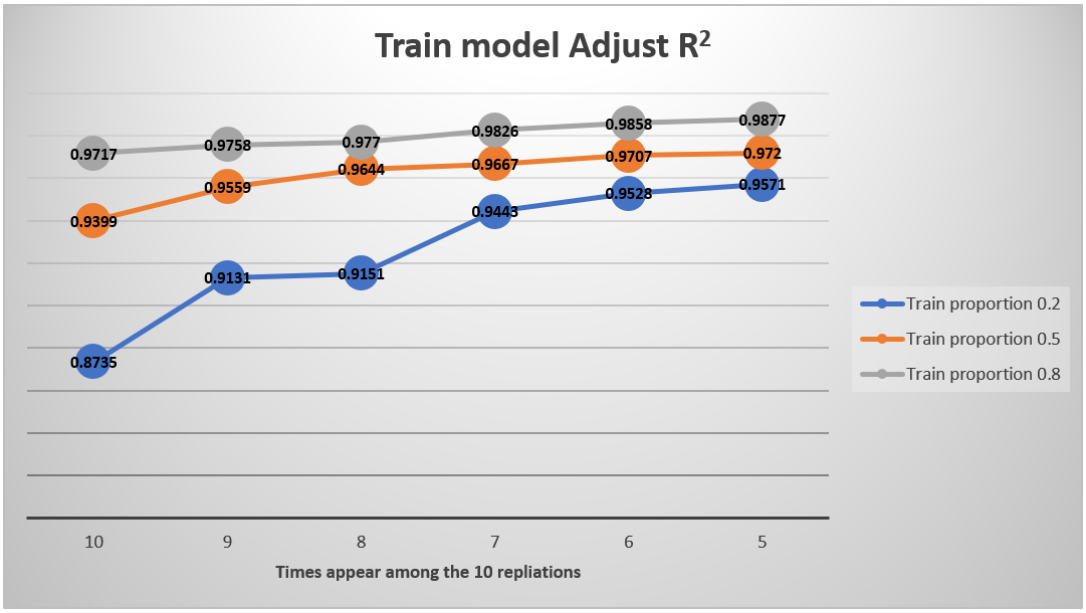

Figure 8. Compare the Train models using different cut-off points.

As expected, both a higher train proportion and reducing the cut-off of the recurrence times will select more CpGs and lead to a model with high adjusted R^2.

Figure 9 shows both a higher train proportion and reducing the cut-off of the recurrence times will train a model has less MSE in test data.

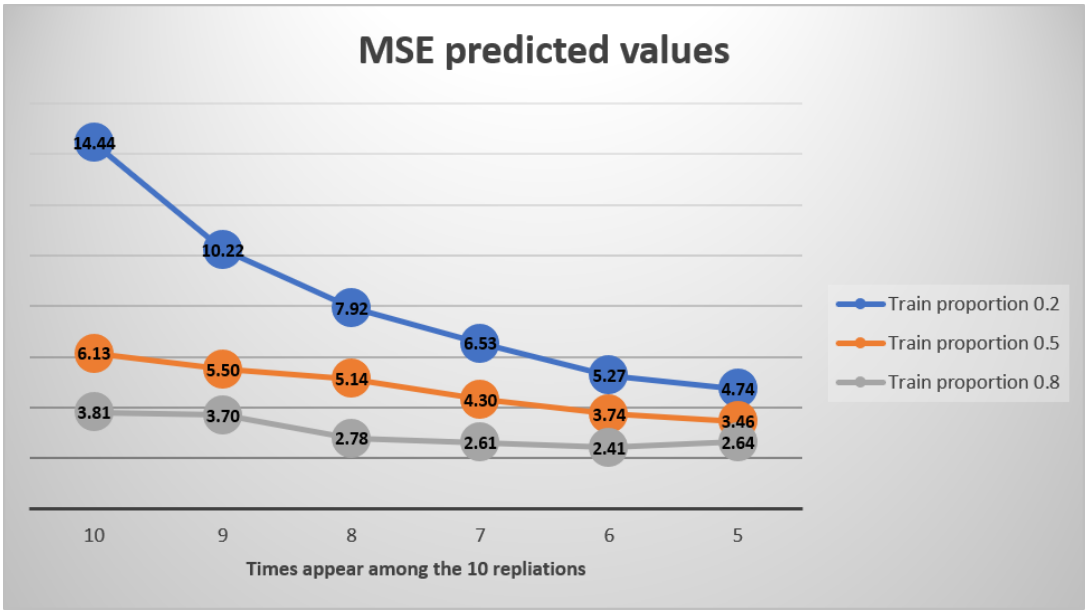

Figure 9. Compare the MSE of the predict values by applying the trained models.

Furthermore, the AIC/BIC of the trained models are showed in table 8. At the recurrence rates 100% and 90%, the sample proportion 0.8 provides the best CpGs to train the model (93 CpGs with $R^2 = 0.9764$ and 151 CpGs with $R^2 = 0.9807$ respectively). 151 CpGs are 1.6 time of 93 CpGs but the R^2 does not improve too much. Thus, in this study, the main models are from the 100% recurrence rate as cut-off. At the recurrence rate 80%, AIC and BIC give different CpGs selection which indicates an unstable result.

Table 8. Compare the AIC/BIC among the trained models.

	AIC	BIC
Overall model		
CPG602	6334.677	9750.533
100% Recurrence rate		
Sample Proportion 0.8 (CPG93)	4349.437	4820.851
Sample Proportion 0.5 (CPG29)	4830.773	4984.602
Sample Proportion 0.2 (CPG7)	5786.153	5830.814
90% Recurrence rate		
Sample Proportion 0.8 (CPG151)	4093.325	4852.548
Sample Proportion 0.5 (CPG49)	4608.031	4861.105
Sample Proportion 0.2 (CPG10)	5583.768	5643.315
80% Recurrence rate		
Sample Proportion 0.8 (CPG206)	4000.398	5032.544
Sample Proportion 0.5 (CPG76)	4347.531	4734.586
Sample Proportion 0.2 (CPG16)	5084.595	5173.915

3.2 Validation test

To validate the performance of our selected CpGs and trained models, the trained models are applied to the validation data (528 patients).

Validation results for 602 CpGs models from overall 2,112 patients' cohort

602 CpGs model is based on elastic net regression of the overall training cohort. In sub-sample elastic regression with recurrence rate 100% as cut-off, 93 CpGs (sample proportion 0.8), 29 CpGs (sample proportion 0.5) and 7 CpGs (sample proportion 0.2) are selected respectively.

In the sub-sample elastic net regression, after selecting the CpGs, the 2,112 patients are split into train and test cohort (0.5 vs. 0.5). The train cohort is used to train the model and the test cohort is used to test the trained model. These models are further validated in the 528 patients independent validation cohort. The MSE/RMSE for train, test and validation data are showed in table 9.

Table 9. MSE/RMSE for models based on different selected CpGs in validation data.

	MSE(RMSE) Train data	MSE(RMSE) Test data	MSE(RMSE) Validation data
602 CpGs	0.66 (0.81)	-	4.16 (2.04)
93 CpGs	3.01 (1.73)	3.41 (1.85)	4.66 (2.16)
29 CpGs	5.35 (2.31)	5.94 (2.44)	6.93 (2.63)
7 CpGs	13.80 (3.71)	14.77 (3.84)	14.21 (3.77)

The MSE/RMSE for 602 CPG model in training data is based on the overall 2,112 patients, so it does not mean too much, as it may overfit the data. By applying the trained model in validation dataset, the RMSE is 2.04 for 602 CpGs model, which means their predict values are ±2.04 years from true values. The 93 and 29 CpGs models can predict patients age with ±2.19 to ±2.60 years from true age. And the core 7 CpGs model can predict patients age with 3.77 years deviant. Comparing to the RMSE of the 602 CpGs models, the RMSEs are a little increasing, but the numbers of CpGs are significantly decreasing.

Figure 10 shows fitted values vs. DNA sampling age in the 528 patients' validation dataset by applying the trained models from 602/93/29/7 CpGs models. The plot shows a good of fit.

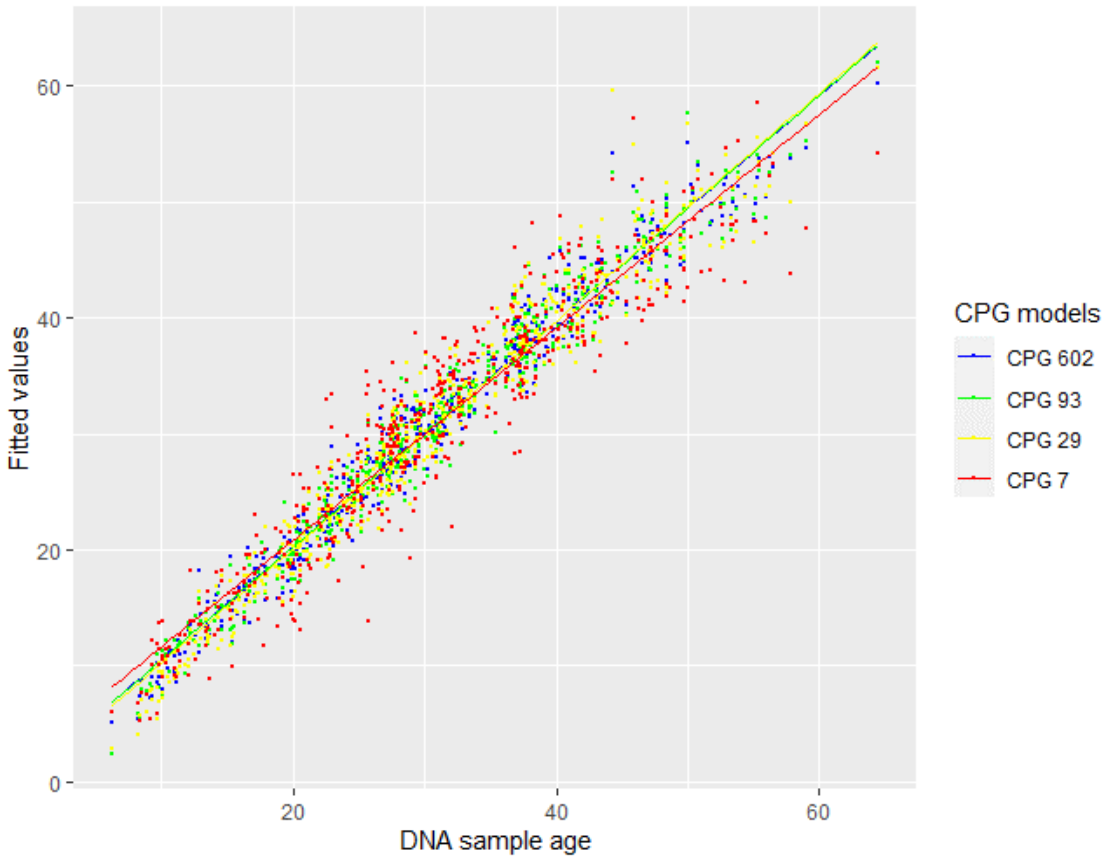

Figure 10. Fitted values vs. DNA sample age using trained CpGs models.

Comparison of the developed models with for published epigenetic clocks

There are 4 published epigenetic clocks: Hannum clock (71 CPG sites), Horvath clock (353 CPG sites), Levine clock (513 CPG sites) and Grimage clock (1030 CPG sites). To show the quality of the selected CpGs and trained models, the MSE/RMSE and beta coefficients are compared with the 4 epigenetic clocks.

It is observed that there exists an overestimate problem in the 4 published epigenetic clocks. An adjustment of $\bar{y} - \hat{\bar{y}}$ is applied to the predicted values from the 4 published epigenetic clocks.

A smaller RMSE and a beta coefficient closer to 1 means the trained model has good efficacy to predict the chronical age.

Table 10. Models comparison of developed models with 4 published epigenetic clocks

	MSE	RMSE	Beta (slope)	Intercept
602 CpGs	4.16	2.04	0.972	0.878
93 CpGs	4.66	2.16	0.973	0.828
29 CpGs	6.93	2.63	0.982	0.464
7 CpGs	14.21	3.77	0.917	2.513
Horvath clock	26.59	5.16	0.636	11.274
Hannum clock	37.52	6.13	0.659	10.542
Levine clock	34.62	5.88	0.860	4.328
Grim age	15.12	3.89	0.788	6.549

In table 10, the RMSEs of the developed models range from 2.04 to 3.77 while the RMSEs from the 4 published clocks range from 3.89 to 6.13. It supports that the developed models have better prediction power than the published 4 aging clocks.

Figure 11 shows the slope (beta estimate) of our models and the 4 published aging clocks. The models we developed have beta estimates ranging from 0.92 to 0.98, which confirms our predicted values are very close to the true chronical age. In the 4 published aging clocks, their beta estimates are less than 0.9 (ranging from 0.64 to 0.86) which means these genetic clocks provide much slower age acceleration estimation. The Levine clock has a beta estimate of 0.86, which is close to our model, but not as good as we have.

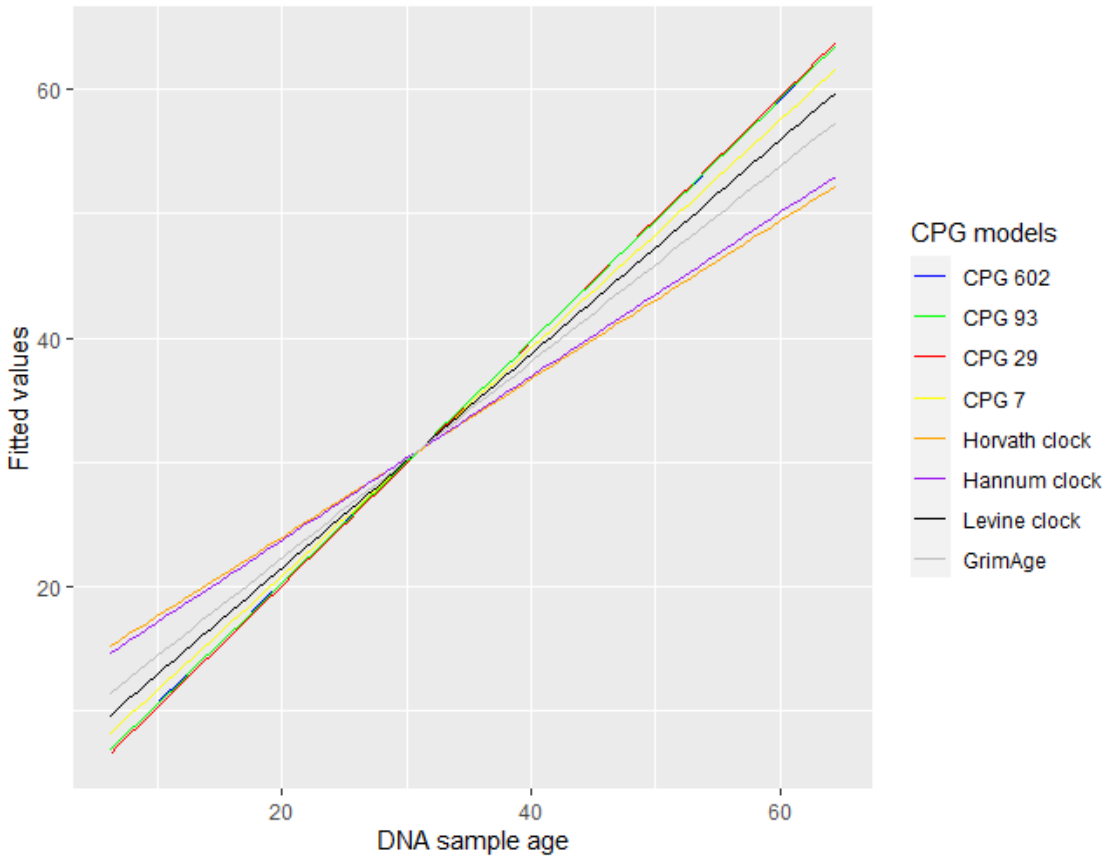

Figure 11. Beta coefficients comparison with the 4 published genetic clocks

3.3 Apply the CpGs models to SJLIFE controls data

To further evaluate the CpGs models we developed, the trained 4 models are applied to SJLIFE control data, which includes 282 independent measured health people. A comparison of the models applied to validation data and control data are showed below:

Table 11. Validate the models in validation(survivor) and control data.

	Validation (Survivor) data				Control data				Survivors vs Controls	Survivors vs Controls
	MSE	RMSE	Beta (slope)	Intercept	MSE	RMSE	Beta (slope)	Intercept	Beta comparison P-value	Intercept comparison P-value
CpG 602	4.16	2.04	0.972	0.878	4.27	2.07	0.922	2.524	0.359	0.022
CpG 93	4.66	2.16	0.973	0.828	4.87	2.21	0.916	2.859	0.344	0.008
CpG 29	6.93	2.63	0.982	0.464	6.12	2.47	0.902	3.232	0.446	<0.0001
CpG 7	14.21	3.77	0.917	2.513	13.85	3.72	0.808	7.352	0.272	<0.0001

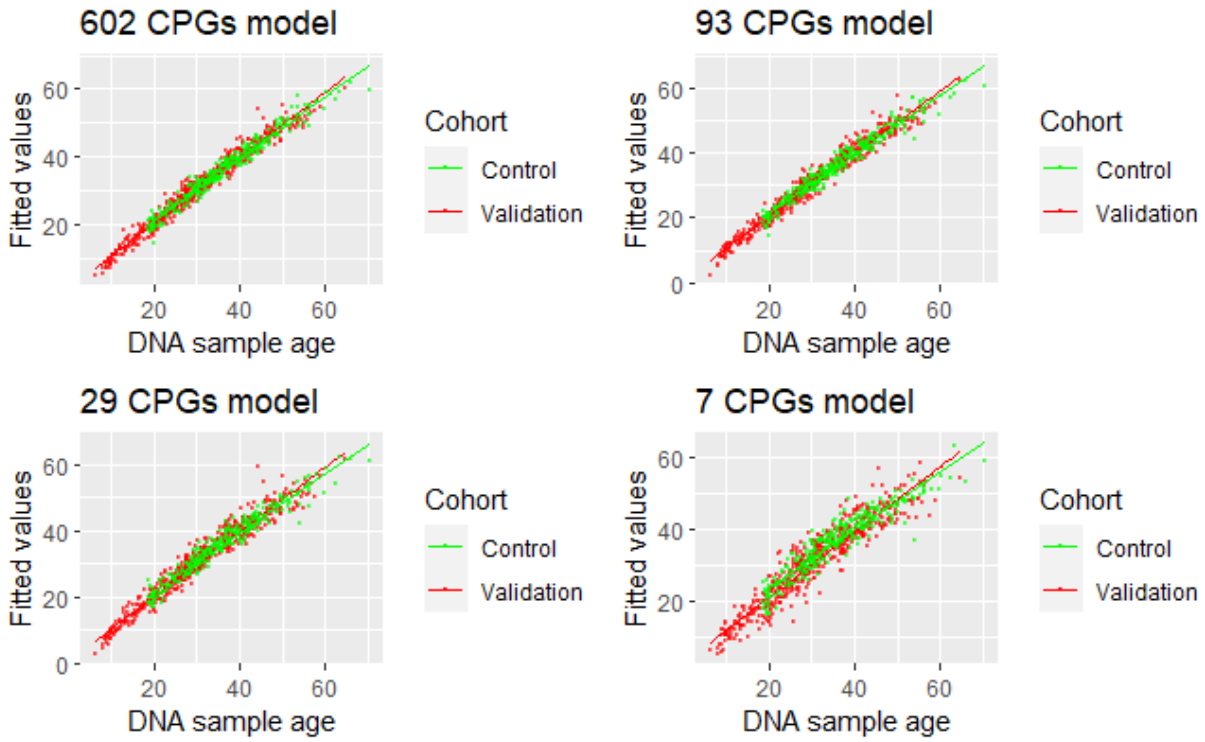

Figure 12. Models validation in validation and control data

The results above tell us the CpGs model we developed performs well in both survivor and control cohort. There is no statistical difference in the beta coefficients among survivor and

controls but the intercept does. This result indicates a potential general using of the developed models regardless their cancer status.

Comparison of the developed models with for published epigenetic clocks in control data

A similar comparison is done with the controls. An adjustment of $\bar{y} - \hat{\bar{y}}$ is applied to the predicted values from the 4 published epigenetic clocks as well.

Table 12. Models comparison of developed models with 4 published epigenetic clocks

	MSE	RMSE	Beta (slope)	Intercept
602 CpGs	4.27	2.07	0.922	2.524
93 CpGs	4.87	2.21	0.916	2.859
29 CpGs	6.12	2.47	0.902	3.232
7 CpGs	13.85	3.72	0.808	7.352
Horvath clock	29.21	5.40	0.586	14.840
Hannum clock	29.77	5.46	0.602	14.263
Levine clock	27.12	5.21	0.742	9.258
Grim age	14.04	3.75	0.784	7.729

In table 12, the RMSEs of the models we developed range from 2.07 to 3.72 while the RMSEs from the 4 published clocks range from 3.75 to 5.46. It supports that developed models have better prediction power than the published 4 aging clocks in controls.

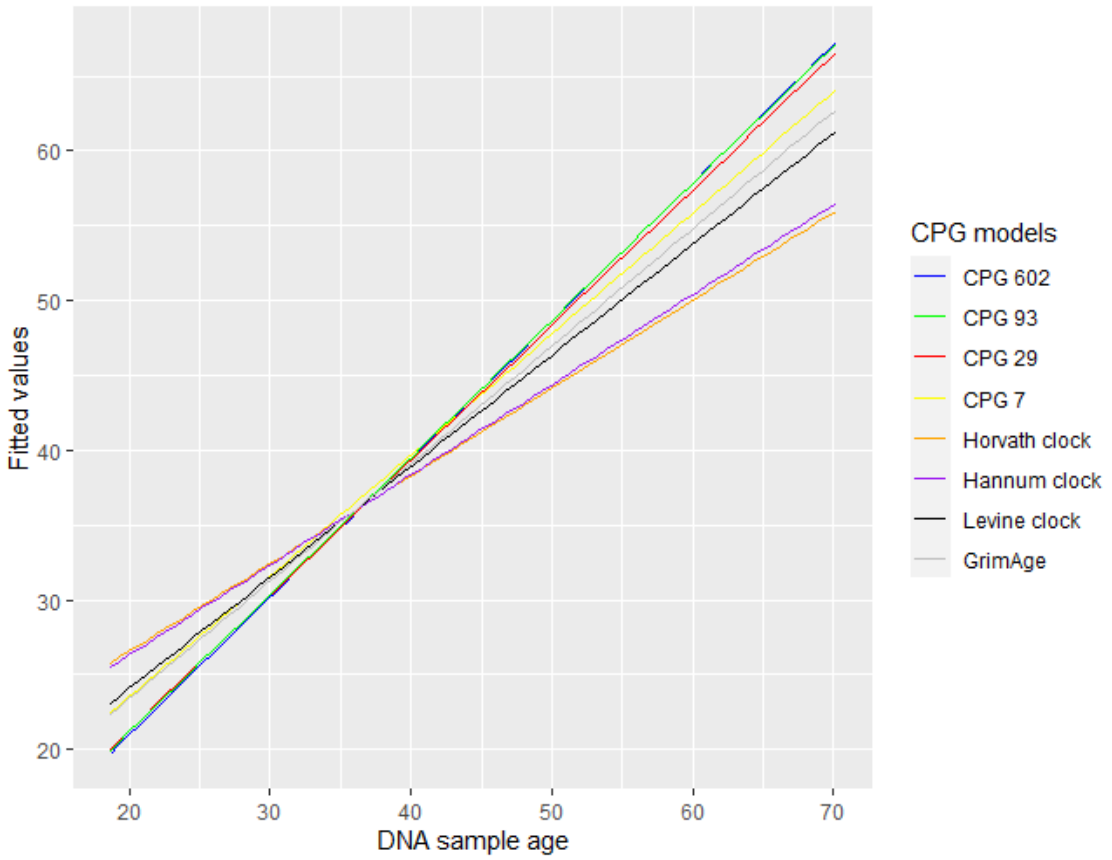

Figure 13. Beta coefficient comparison with the 4 published genetic clocks in controls

Figure 13 shows the slope (beta estimate) of our models and the 4 published aging clocks in controls. The models we developed have beta estimates ranging from 0.81 to 0.92, which confirms our predicted values are very close to the true chronical age. In the 4 published aging clocks, their beta estimates are less than 0.8 (ranging from 0.59 to 0.78), which means these genetic clocks provide much slower age acceleration estimation.

Although, the published epigenetic clocks showed a little off in their performance in chronical age prediction, our developed SJLIFE clocks showed strong correlation ($r < 0.8$) with the published clocks of their performance (figure 14-17).

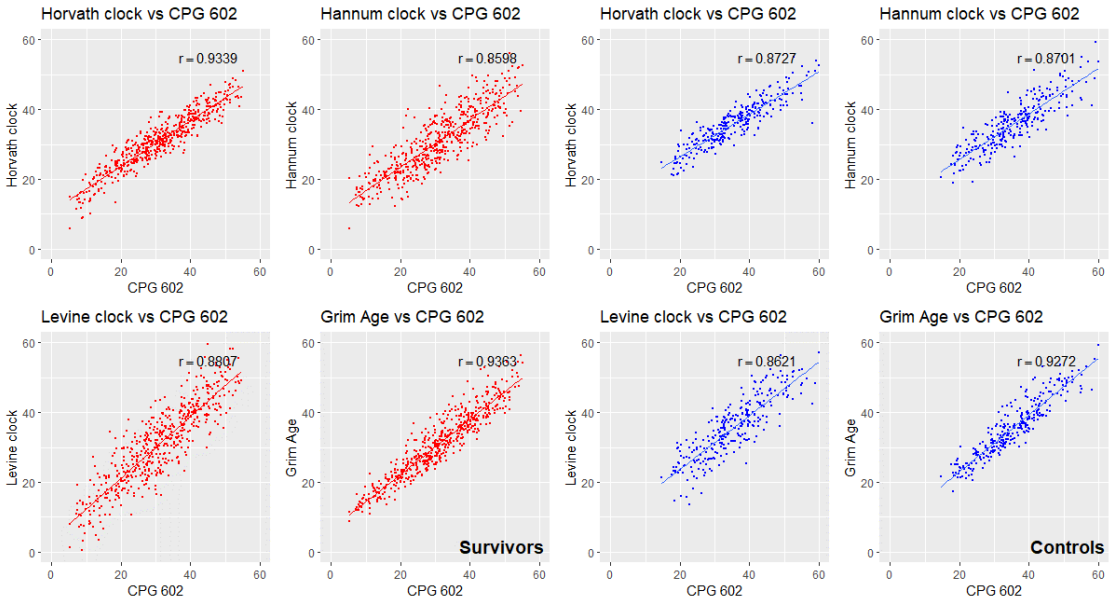

Figure 14. CPG602 vs four published clocks in chronical age prediction

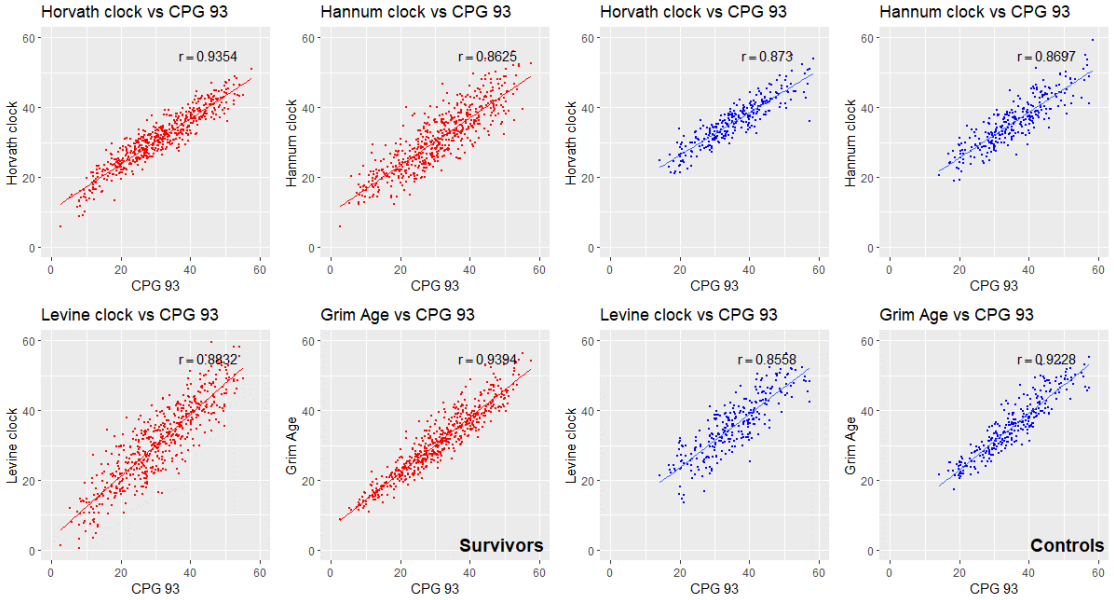

Figure 15. CPG93 vs four published clocks in chronical age prediction

Figure 16. CPG29 vs four published clocks in chronical age prediction

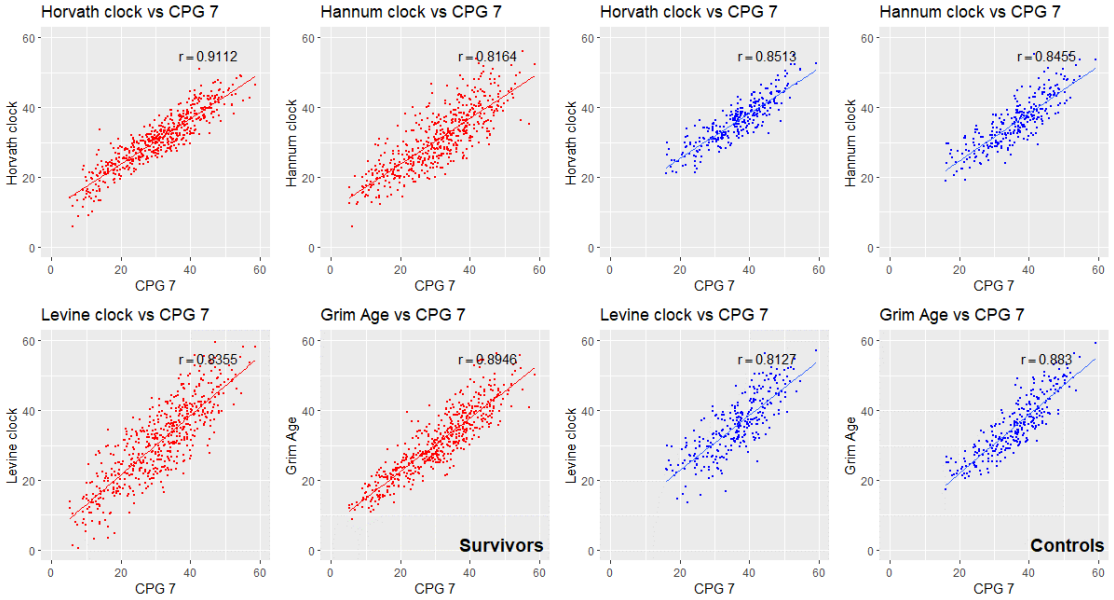

Figure 17. CPG7 vs four published clocks in chronical age prediction

3.4 Summary of variables selection using elastic net regression

In ultra-high dimension data analysis, the elastic net regression performs well to select the initial candidate variables, though the over-fitting problems may exist. The sub-sampling and ensemble method provides a way to further reduce the selected features by setting different sample proportion and inclusion criteria. It is proven that the core variables will be selected with smaller training cohorts and higher recurrence rates.

Using the proposed method in SJLIFE epigenetic colock study (2,112 patients training cohort), 602 CpGs are selected as the best CpGs set among the 6 random split cohort in selection/training process. The 602 CpGs produce a good prediction model with adjusted R^2 of 0.9928. And the trained model performs well in the validation cohort and control cohort with RMSEs 2.04 and 2.07 respectively.

The recurrent sub-sample elastic net regression technique can select core features. In the CpGs analyses, 93, 29, 7 CpGs are selected for sample proportion 0.8, 0.5 and 0.2 respectively. And corresponding trained models perform well in both validation and control cohort as well as outperforms the 4 published epigenetic clocks.

In short, from the original 689,414 CpGs we selected 602 CpGs to build the aging clock for SJLIFE cancer survivors and controls with great performance. And we further reduce it to 7 core CpGs which is can explain about 89% variation of the chronical age.

Chapter 4

New variable selection procedures for big data

4.1 Potential issue with ultra-high dimension data

The elastic net regularized regression is a very powerful and widely used variable selection method in high dimension data analysis as long as the data can be fitted into the memory. It usually requires large memory on HPCs. But with rapid growth of big data volumes in both n (observation numbers) and p (parameter numbers), sometimes the machine's memory is not big enough to handle the data. In this case, the regular elastic net regression cannot be directly applied to ultra-high dimension data.

To solve the memory problem, we may need a more powerful HPC, but may cost more. An alternative way is to reduce the memory use in each iteration of the analysis. In our study, we propose the partition elastic net regression method which splits the big data into smaller datasets and then apply the regular elastic net method to each of the sub datasets in order to reduce the memory requirement. Then further analyses will be performed based on the combined results to make the ultra-big data analysis feasible. In our design, there are 3 ways to split the data: random split the by rows if the number of observation numbers (n) is large, random split the data by columns if the number of parameters (p) is large, and random split the data by blocks if both n and p are large.

The reason that we use the random partition instead of random sampling is that random partition can keep all the subjects or parameters in the sub-dataset. But the random sampling method may loss some important observation or variables.

4.2 Partition by rows

General random row partition procedure

Let $D = (X, y)$ be the data matrix with y be the n-dimension vector corresponding to response and X be p input variables of size $n \times p$. Let R be the number of (random) row partitions

of D into, say, $D_r = (X_r, y_r)$, where $r = 1, 2, \cdots, R$ with approximately the number of rows, n/R.

Let S be the number of random selections/ensembles needed and s be its running index for each iteration, $s = 1, 2, \cdots, S$ below.

- For $s = 1, 2, \cdots, S$, do
  - For each $r = 1, 2, \cdots, R$, perform variable selection (Elastic net) procedure on each subdata $D_r = (X_r, y_r)$ to select the reduced variable set, say, $T_{r,s}$.
- Ensemble these variable by taking union of set $T_{r,s}$

$$\mathrm{CPG}_{ROW} = \cup_{s=1}^{S} \cup_{r=1}^{R} T_{r,s}.$$

- Use this set to select the new data matrix to replace original data matrix $D = X$ for subsequent analysis.

$$D_{ROW} = (XC_{ROW}, y),$$

where C_{ROW} is the binary matrix used to select the columns of X according to reduced variable set CPG_{ROW}.

4.3 Row partition elastic method application

We apply the above general row-partition procedure for our CpG selection with $n = 2,112$ and $p = 689,414$ with $R = 2$ and $S = 10$. Our results show that it greatly reduce the number of columns from $689,414$ to $5,237$ with comparable performance on subsequent analysis. For example, we can use the reduce matrix of size $n \times 5,237$ for additional further variable reduction on the CpGs vs various sub-sample procedure as discussed previously.

Selected CpGs and compare with the reference CpG set (602 CpGs)

A second regular elastic net analysis is applied to the 5,237 CpGs. And 585 CpGs are selected. Comparing these CpGs with the reference CpG set (602 CpGs), there are 564 CpGs are in common (93.69% of the 602 CpGs). It indicates that the partition elastic net regression by rows

is a potential eligible replacement for regular elastic net method. Figure 18 shows the overlap between the 602 CpGs (reference set) and the 585 CpGs (partition EN by rows).

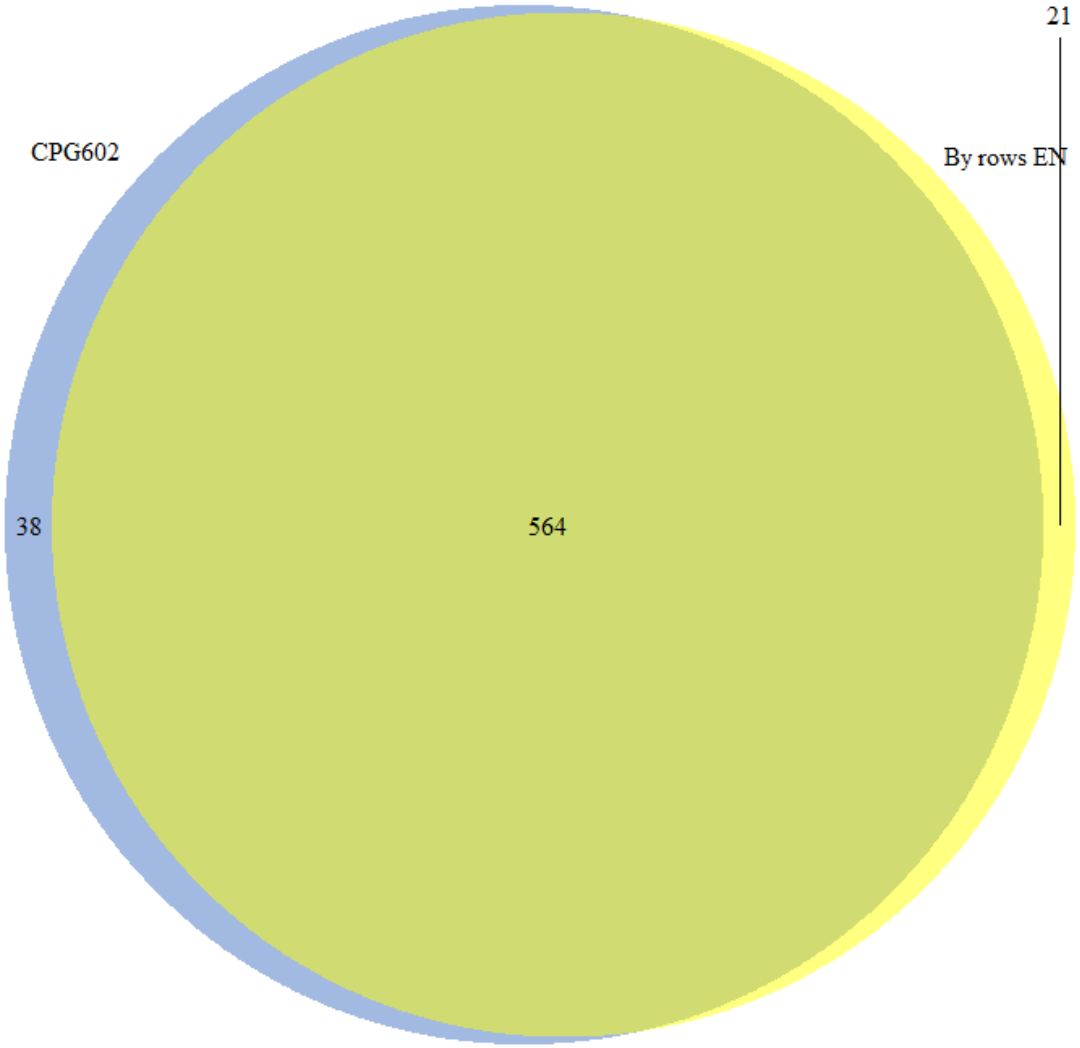

Figure 18. Comparison for the selected 585 CpGs (partitions by rows) with the reference set (602 CpGs).

4.4 Partition by columns

General random column partition procedure

Let $D = (X, y)$ be the data matrix with y be the n-dimension vector corresponding to

response and X be p input variables of size $n \times p$. Let C be the number of (random) columns partitions of X into, say, $X = (X_1, X_2, \cdots, X_C)$, X_c is its c-th where $c = 1, 2, \cdots, C$ with approximately the number of columns, p/C, and set $D_c = (X_c, y)$.

Let S be the number of random selections/ensembles needed and s be its running index for each iteration, $s = 1, 2, \cdots, S$ below.

- For $s = 1, 2, \cdots, S$, do
  - For each $c = 1, 2, \cdots, C$, perform variable selection (Elastic net) procedure on each subdata $D_c = (X_c, y)$ to select the reduced variable set, say, $T_{c,s}$.
- Ensemble these variable by taking union of set $T_{c,s}$

$$\mathrm{CPG}_{COL} = \cup_{s=1}^{S} \cup_{c=1}^{C} T_{c,s}.$$

- Use this set to select the new data matrix to replace original data matrix $D = X$ for subsequent analysis.

$$D_{COL} = (XC_{COL}, y),$$

where C_{COL} is the binary matrix used to select the columns of X according to reduced variable set CPG_{COL}.

4.5 Column partition elastic net method application

We apply the above general column-partition procedure for our CpG selection with $n = 2,112$ and $p = 689,414$ with $C = 2$ and $S = 10$. Our results show that it greatly reduce the number of columns from $689,414$ to $4,012$ with comparable performance on subsequent analysis, such as various sub-sampling procedures, with the reduce matrix of size $n \times 4,012$.

Selected CpGs and compare with the reference CpG set (602 CpGs)

A further regular elastic regression selected 602 CpGs from the 4,012 CpGs pool. Comparing these CpGs with the reference CpGs (602 CpGs), there are 599 CpGs are in common (99.50% of the 602 CpGs). It indicates that the partition elastic net regression by columns is a very good

alternative method to selected CpGs in case the memory is not enough to perform the regular elastic net regression. Figure 19 shows the overlap between the reference 602 CpGs and the 602 CpGs selected from the partition elastic net by columns.

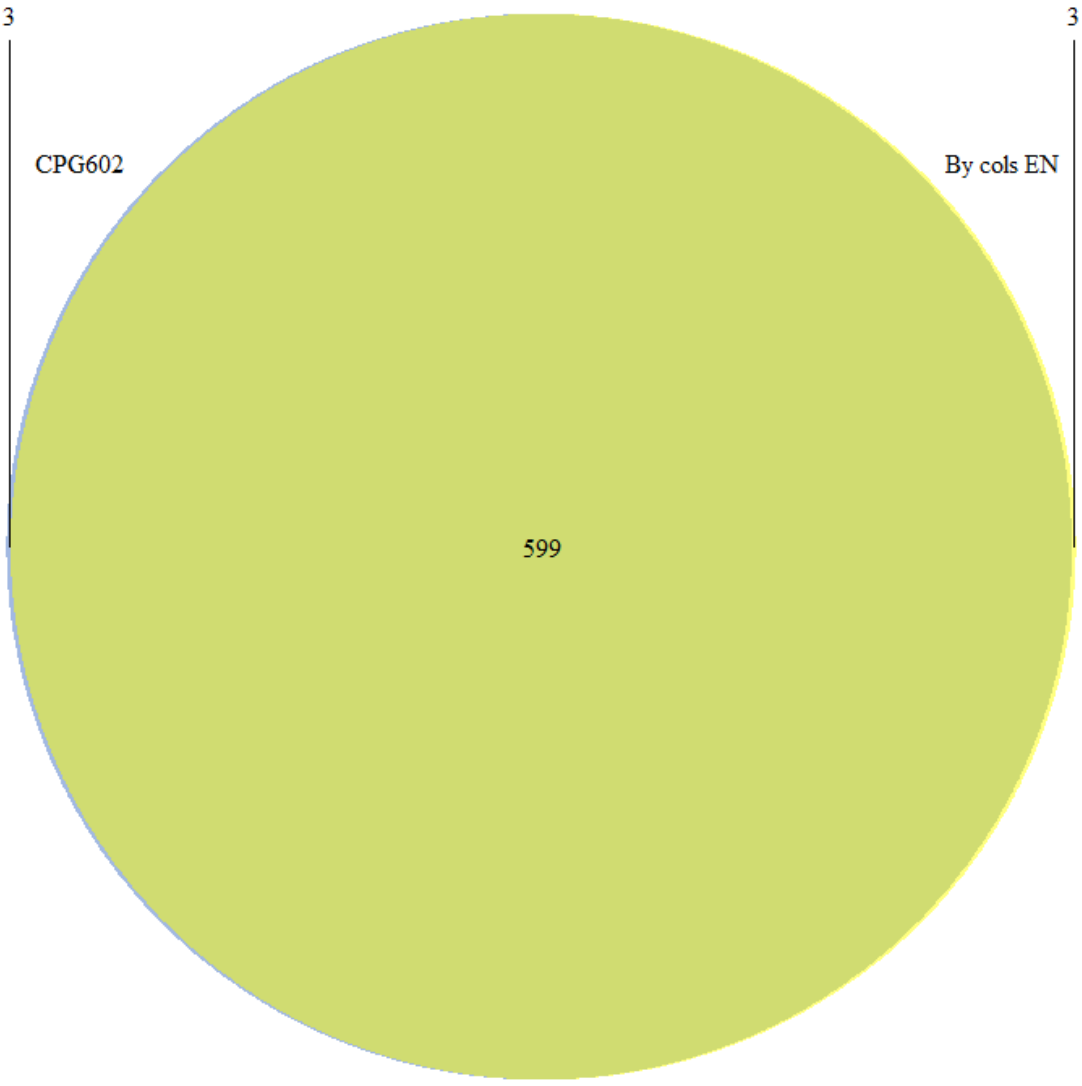

Figure 19. Comparison for the selected 602 CpGs (partitions by columns) with the reference CpG set (602 CpGs).

A further comparison of column partition method and random sample by column method

In the above section, the column partition elastic net method showed almost a perfect result in CpGs selection as an alternative way of regular elastic net regression. In this section we want to compare the performance in CpG selection between the column partition method and the random sample by column method. The major difference between these two methods is the column partition method performs the elastic net regression on each of the split part while the random sample method only analyzes the selected sub-sample data. In this case, if the iteration numbers are small each columns/variables will be analyzed in column partition methods while the random sample by column method will skip many of the columns/variables. But we can expect with the number of iterations increasing, there will be less of a difference between the column partition elastic net regression and the random sample by column partition elastic net regression. To show this, we first random split the overall CpGs data into half vs. half part and perform the elastic net regression on each part with 5 iterations. Similarly, we random select two samples with replacement with sample proportion 0.5 and perform the elastic net regression on each sample with 5 iterations. Then, basing on the selected CpGs from each iteration, CpGs pools are created from different iteration numbers (eg. Union CpGs for 1 iteration, 2 iterations, 3 iterations etc.) for column partition method and random sample by column method respectively. The number of CpGs in these CpGs are much smaller compared to the original number of CpGs (689,414). Next, the regular elastic net regression is performed on these CpGs pool and final CpGs sets are selected. Table 13 shows the number of selected CpGs in the reference GpGs (602 CpGs). It is clear that when the iteration numbers are small the column partition method are more efficient in CpGs/variables selection. With the iteration numbers increasing, the random sample by column method catch up very quickly.

Table 13. Comparison of the selected reference CpGs between column partition elastic net method and random sample by column elastic net method with different iteration numbers.

Number of iterations	Column partition	Random sample by column
1 iteration	435	336
2 iterations	473	411
3 iterations	574	556
4 iterations	576	569
5 iterations	578	580

The figure 20 visualizes our findings in table 18.

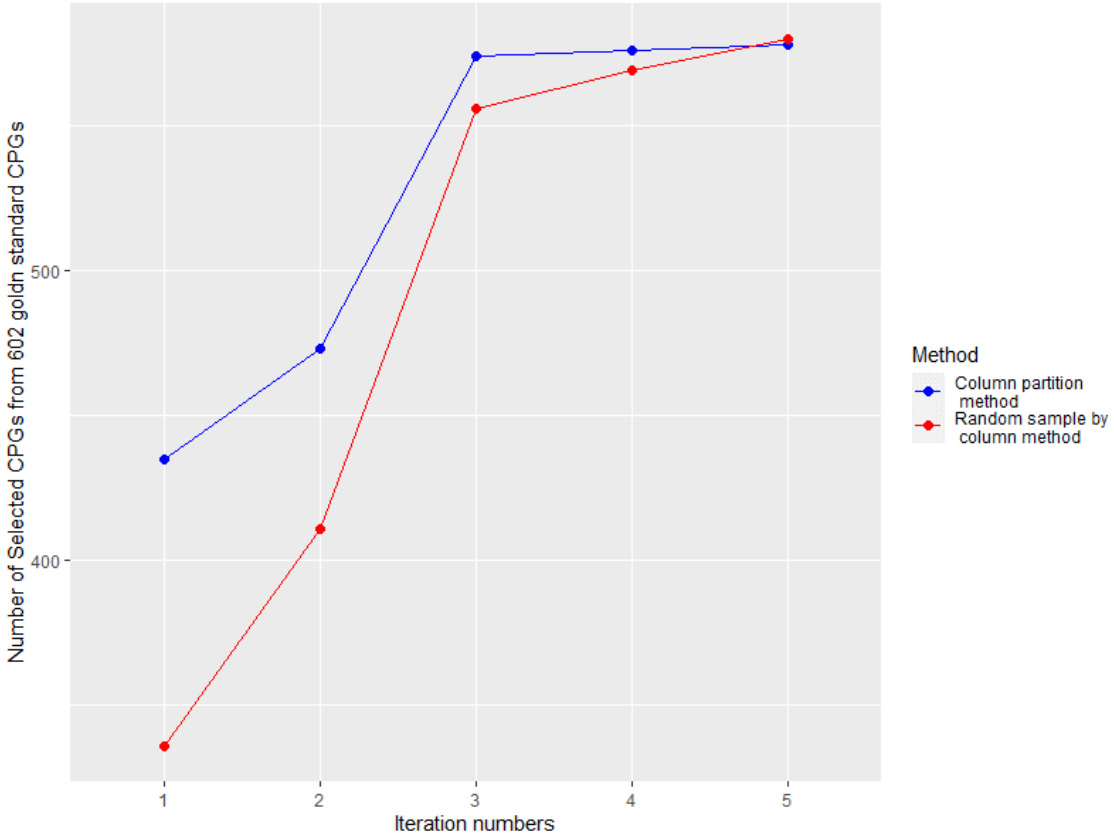

Figure 20. Comparison of the selected reference CpGs between column partition elastic net method and random sample by column elastic net method with different iteration numbers.

4.6 Partition by blocks

General random block/matrix partition procedure

Let $D = (X, y)$ be the data matrix with y be the n-dimension vector corresponding to response and X be p input variables of size $n \times p$. Let R be the number of (random) row partitions and C be the number of (random) columns partitions of X into $R \times C$ sub-blocks, say,

$$\{(X_{r,c}), r = 1, 2, \cdots, R, c = 1, 2, \cdots, C\}$$

each with approximately the number of rows, n/R, and approximately the same number of columns,

p/C. For each r, we can find the corresponding row partition of y denoted as y_r. Finally, we define the sub-data matrix as $D_{r,c} = (X_{r,c}, y_r)$.

Let S be the number of random selections/ensembles needed and s be its running index for each iteration, $s = 1, 2, \cdots, S$ below.

- For $s = 1, 2, \cdots, S$, do
  - For each $r = 1, 2, \cdots, R$ and each $c = 1, 2, \cdots, C$, perform variable selection (Elastic net) procedure on each subdata $D_{r,c} = (X_{r,c}, y_r)$ to select the reduced variable set, say, $T_{r,c,s}$.
- Ensemble these variable by taking union of set $T_{r,c,s}$

$$\mathrm{CPG}_{BLOCK} = \cup_{s=1}^{S} \cup_{r=1}^{R} \cup_{c=1}^{C} T_{r,c,s}.$$

- Use this set to select the new data matrix to replace original data matrix $D = X$ for subsequent analysis.

$$D_{BLOCK} = (XC_{BLOCK}, y),$$

where C_{BLOCK} is the binary matrix used to select the columns of X according to reduced variable set CPG_{BLOCK}.

4.7 Block partition elastic net method application

We apply the above general block-partition procedure for our CpG selection with $n = 2,112$ and $p = 689,414$ with $R = 2$, $C = 2$ and $S = 10$. Our results show that it greatly reduce the number of columns from $689,414$ to $10,945$ with comparable performance on subsequent subsampling analysis with the reduce matrix of size $n \times 10,945$.

Selected CpGs and compare with the reference CpG set (602 CpGs)

By performing a second elastic net regularized regression on the selected 10,945 CpGs, a final list of 586 CpGs are picked out. Comparing these CpGs with the reference CpGs (602 CpGs), there are 571 CpGs are in common (94.85% of the 602 CpGs). It indicates that the partition elastic

net regression by blocks performs well if regular elastic method is not executable. Figure 21 shows the overlap between the 602 CpGs (reference CpGs) and the 586 CpGs (partition EN by blocks).

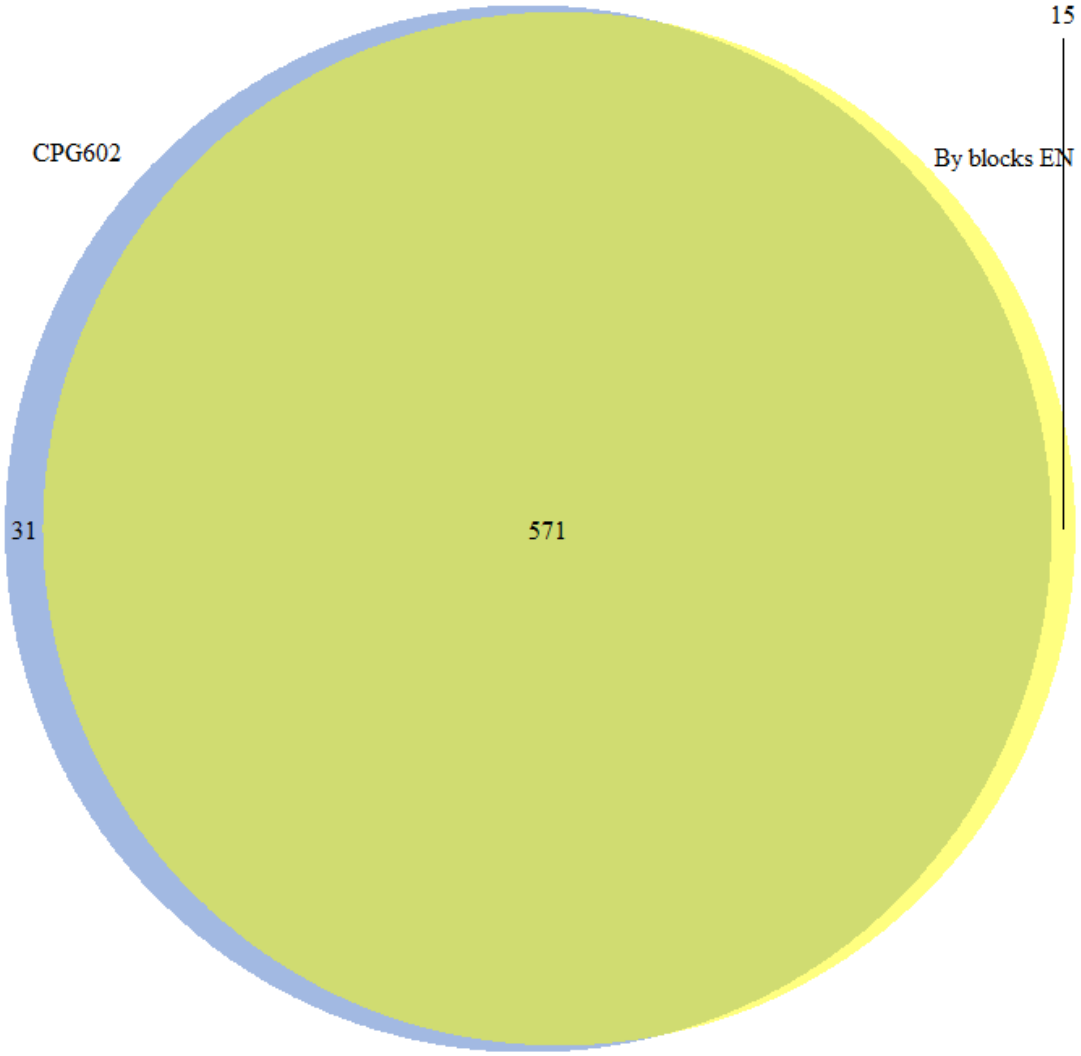

Figure 21. Comparison for the selected 586 CpGs (partitions by blocks) with the reference CpGs (602 CpGs).

4.8 Overall comparison of the different partition methods

In the above analysis, it is showed all the three partition elastic net methods perform well to substitute the regular elastic net regression. Especially, partition elastic net regression by columns

shows almost perfect performance which captures 99.50% CpGs from the reference 602 CpGs. And the Venn-diagram in Figure 22 shows that there are 541 CpGs (89.87% of the 602 CpGs) are in common in the reference 602 CpGs and the other 3 partition elastic net methods.

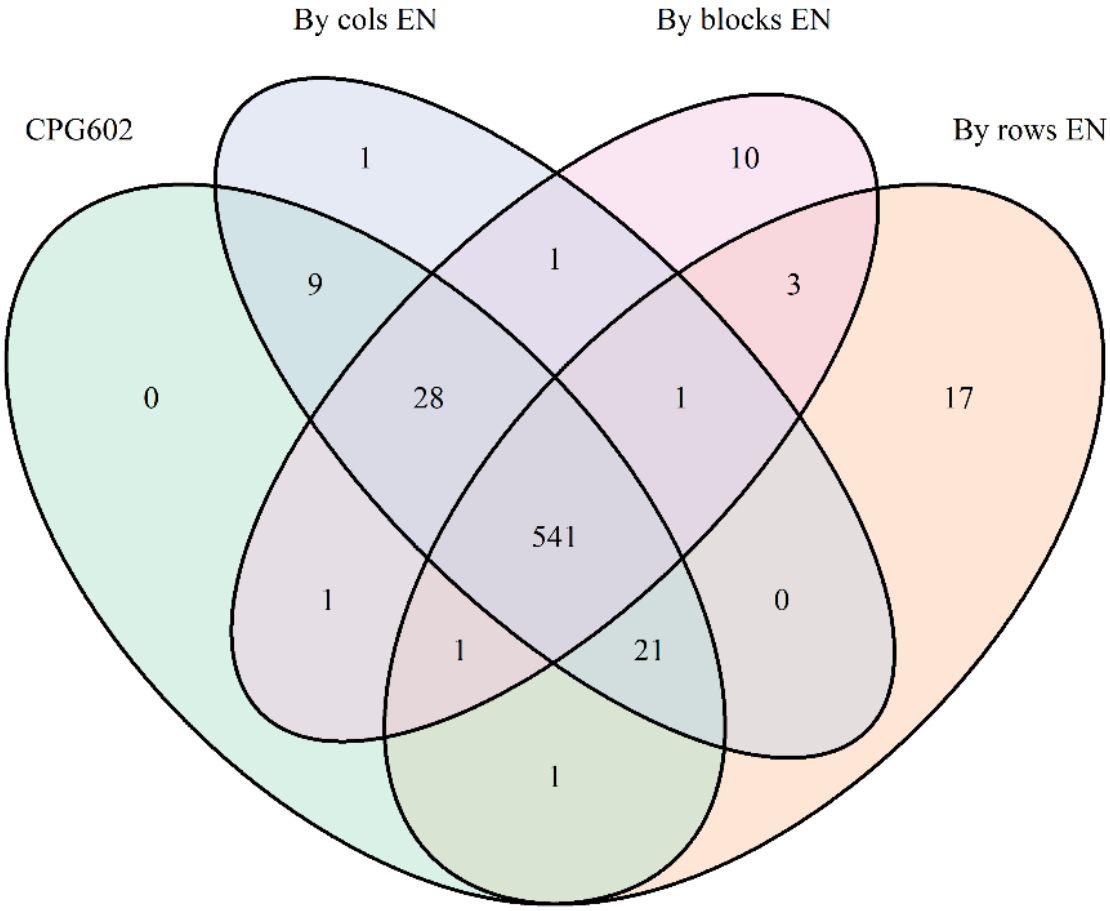

Figure 22. Overall comparison of the reference 602 CpGs and the other 3 partition elastic net methods.

Table 14 shows the adjusted R^2 of the trained models using the selected CpGs from reference set and partition elastic net methods. All of their adjusted R^2 are greater than 0.99 which

indicate the partition elastic net method is good for using when short of memory to perform the regular elastic net regression on cluster.

Table 14. Model comparison of the overall cohort.

	$AdjustedR^2$
CpG 602 (reference set)	0.9928498
CPG 585 (partition by rows)	0.9924814
CPG 602 (partition by columns)	0.9927094
CPG 586 (partition by blocks)	0.9925942

Figure 23 shows the how the predict age fit with the true chronical age using different models trained from the CpGs sets selected from different partition elastic net methods and compare them with the reference set. The plot shows the good of fit of these models and confirms the good performance of the CpGs selection procedure using partition elastic net methods.

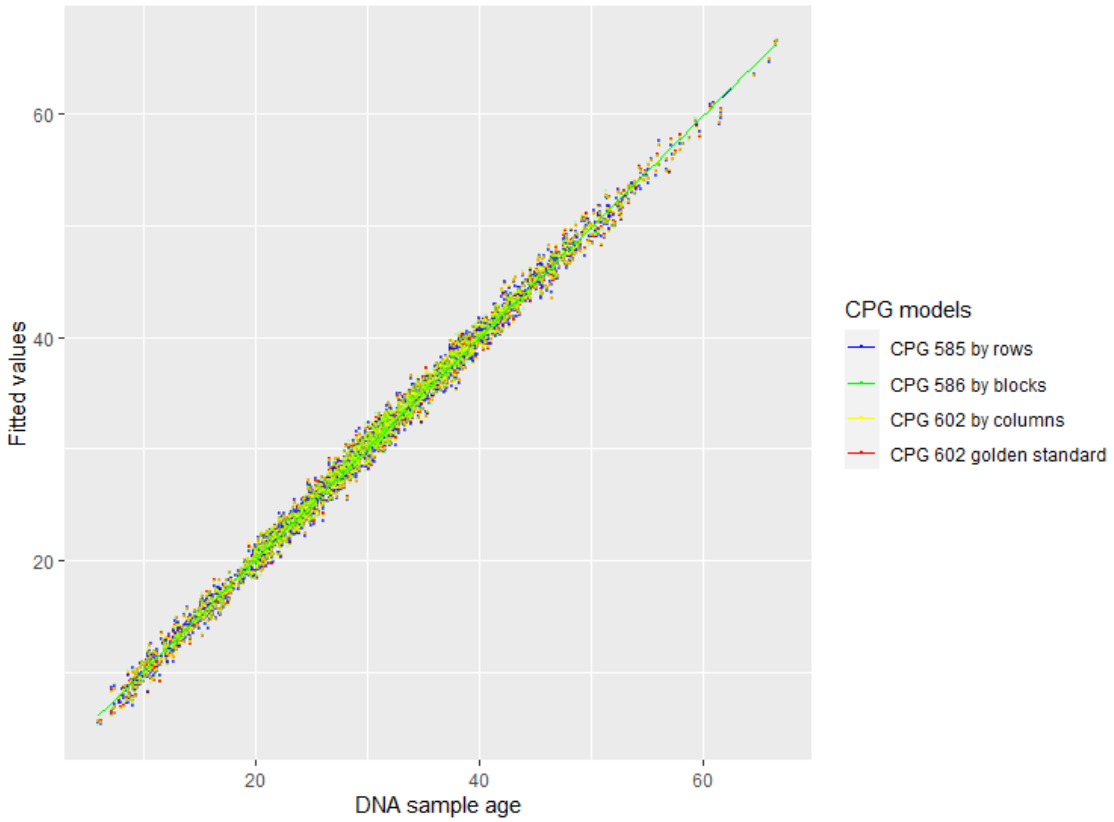

Figure 23. Fitted values vs. DNA sample age using different partition elastic net methods.

4.9 Combination of partition and sub-sample elastic net methods

The sub-sample elastic net regression showed it is efficient in core CpGs selection. We further exploration it's application combining with the partition elastic net methods.

To combine the sub-sample method and partition method together, firstly the partition elastic net method is used to select CpGs pools. Then repeatedly use the sub-sample elastic net method to select the core CpGs.

Sub-sample proportion 0.8

i. By rows:

As shown before, the row partition elastic net method created a CpGs pool with 5,237 CpGs. Then the sub-sample elastic net regression is applied to the 5,237 CpGs with 10 iterations. And 98 CpGs are selected in the final list (100% recurrence rate). Comparing these CpGs with the

reference set (93 CpGs), there are 69 CpGs are in common (74.19% of the 93 CpGs). Figure 24 shows the overlap between the 93 CpGs (reference) and the 98 CpGs (partition EN by rows).

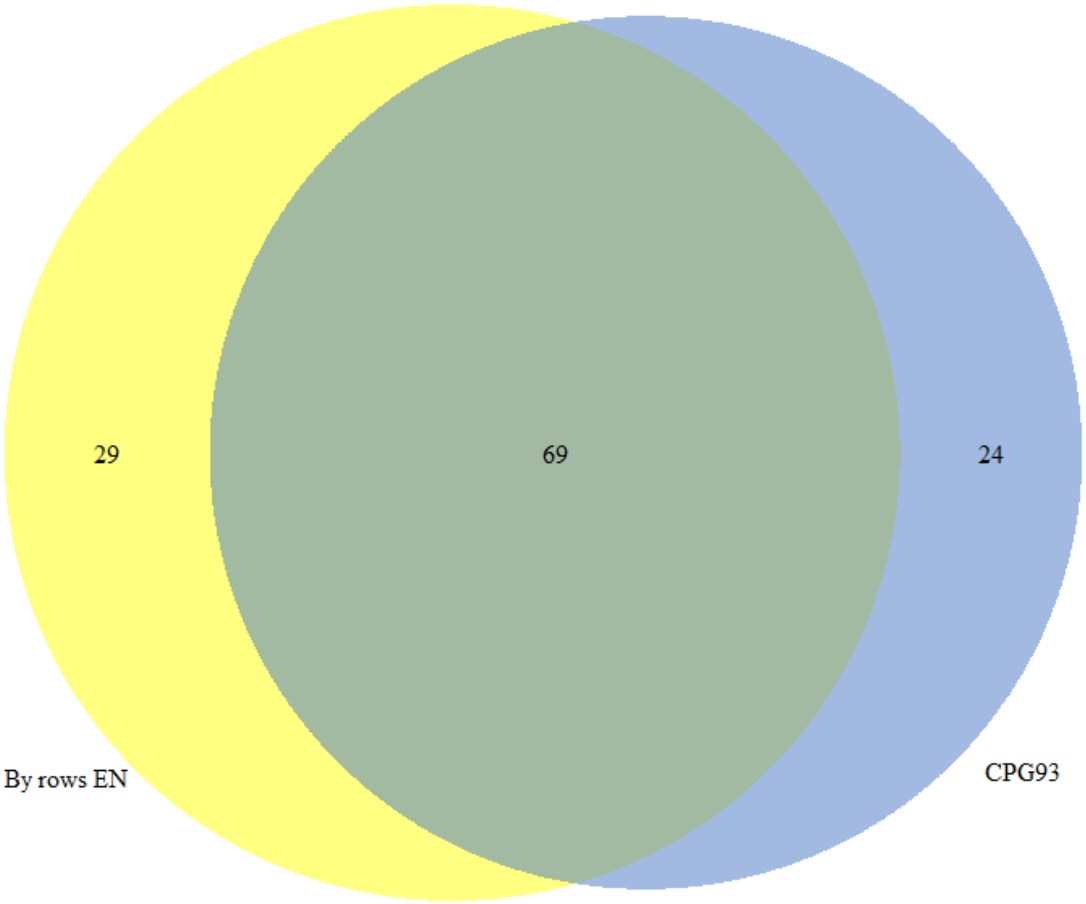

Figure 24. Comparison for the selected 98 CpGs (partitions by rows) with the reference CpGs (93 CpGs).

ii. By columns:

Column partition elastic net method selected 4,012 CpGs pool. And 94 CpGs are selected (100% recurrence rate) after the 10 iterations sub-sample analyses. Comparing the 94 CpGs with

the reference CpGs (93 CpGs), there are 68 CpGs are in common (73.12% of the 93 CpGs). Figure 25 shows the overlap between the 93 CpGs (reference) and the 94 CpGs (partition EN by columns).

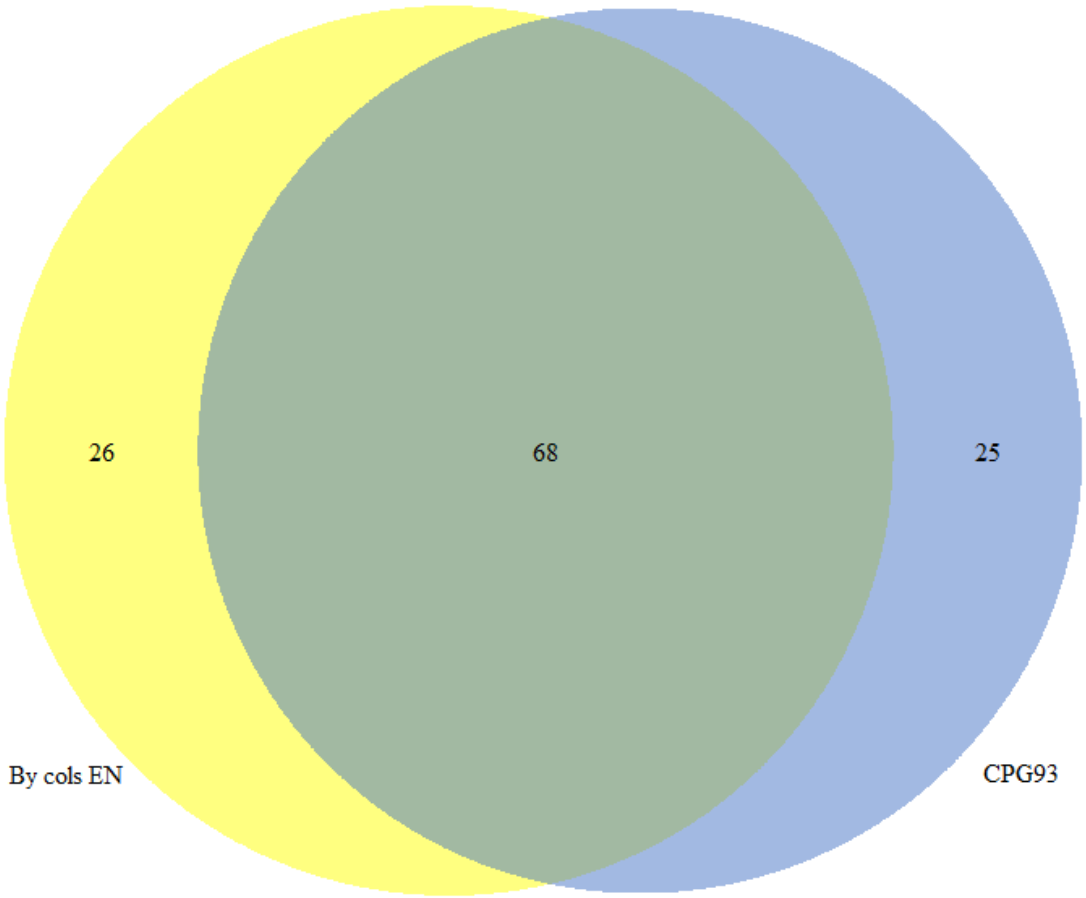

Figure 25. Comparison for the selected 94 CpGs (partitions by columns) with the reference CpGs (93 CpGs).

iii. By blocks:

Block partition elastic net method created a 10,945 CpGs pool. And the 10 iterations sub-sample elastic regression method selected 93 CpGs (100% recurrence rate). Comparing the

selected 93 CpGs with the reference set (93 CpGs), there are 67 CpGs are in common (72.04% of the 93 CpGs). Figure 26 shows the overlap between the 93 CpGs (reference) and the 93 CpGs (partition EN by blocks).

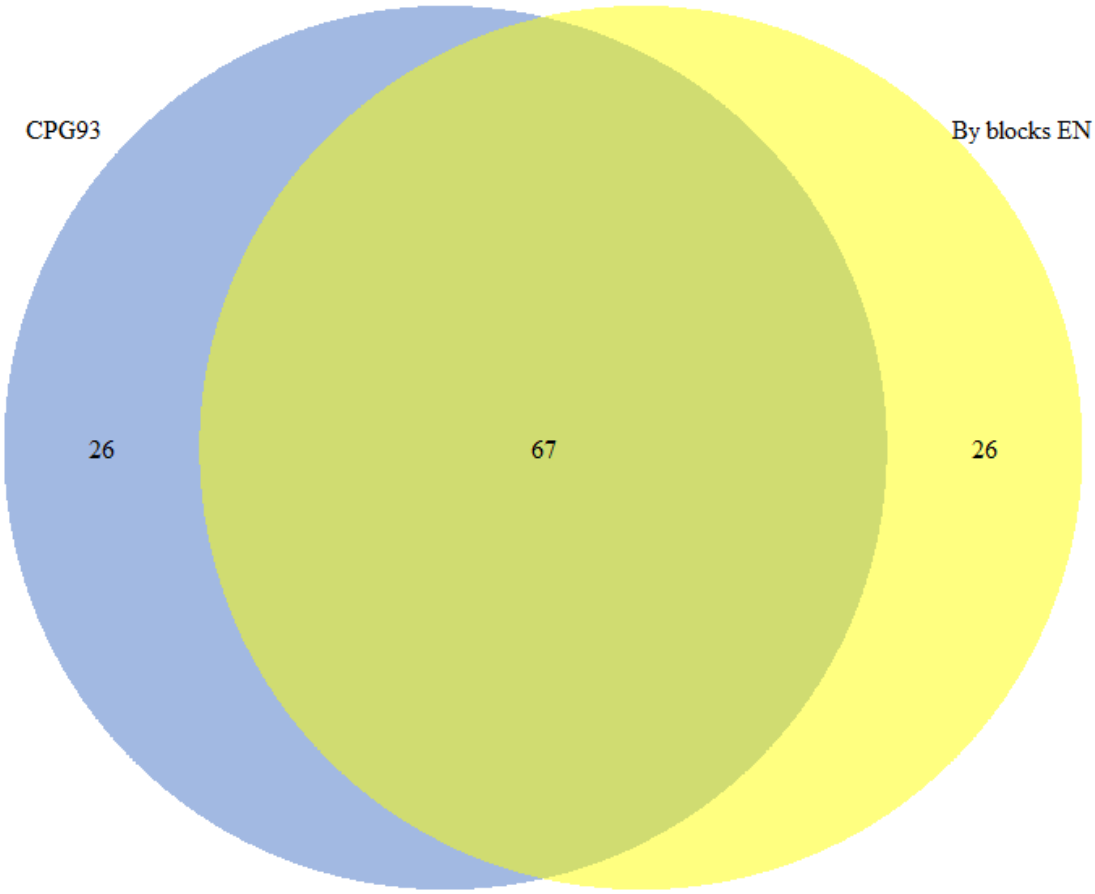

Figure 26. Comparison for the selected 93 CpGs (partitions by blocks) with the reference CpGs (93 CpGs).

The above analysis showed all the three partition elastic net methods perform well when combined with the sub-sample elastic net regression. The results are stable in which 67 to 69 core

CpGs are selected. And the Venn-diagram in Figure 27 shows that there are 56 CpGs (60.2% of the 93 CpGs) are in common in the reference 93 CpGs and the other 3 partition elastic net methods combing with sub-sample elastic net regression with sample proportion 0.8.

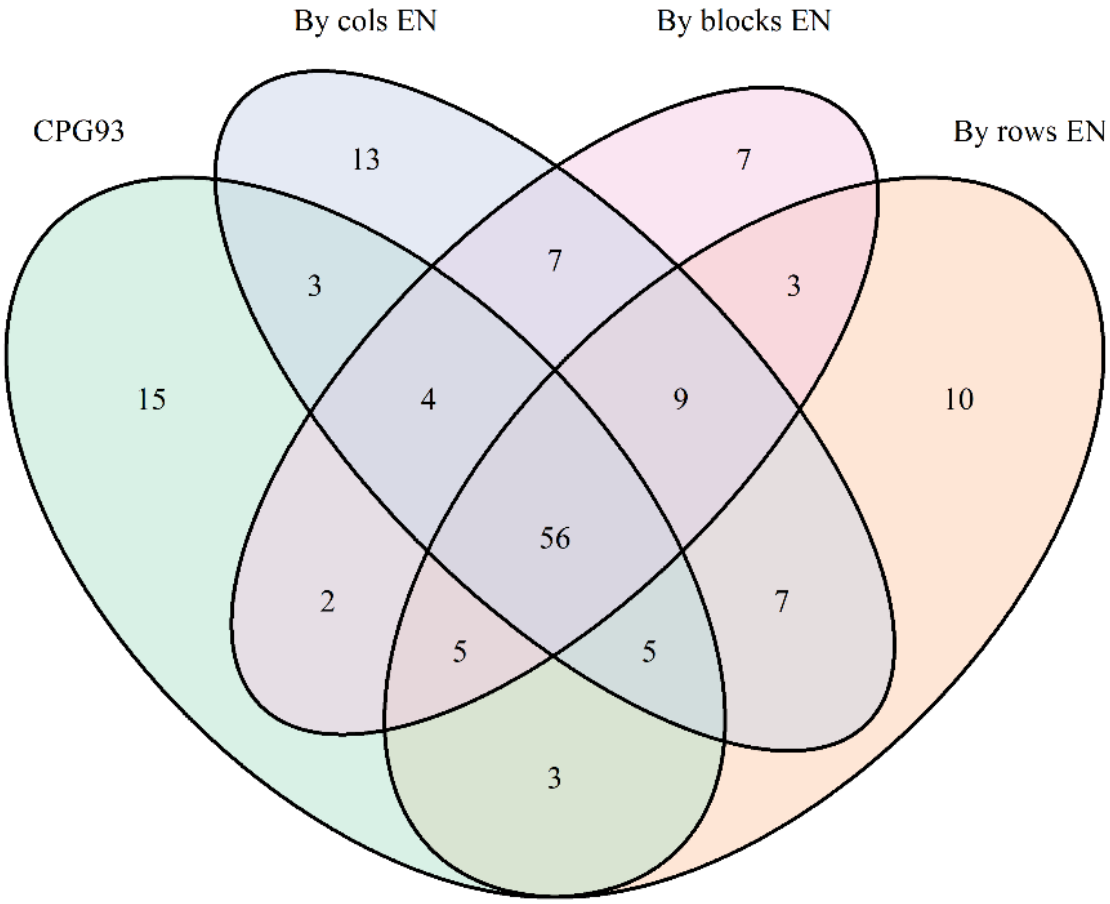

Figure 27. Relationship of the reference 93 CpGs and the CpGs from 3 partition methods with sub-sample proportion 0.8.

Table 15 shows the adjusted R^2 of the trained models using the selected CpGs from the reference set and combining the partition methods and sub-sample elastic net methods with pro-

portion 0.8. All of their adjusted R^2 are greater than 0.97 which indicate the combination of the partition method and sub-sample elastic net regression is good for using when short of memory to perform the regular sub-sample elastic net regression on cluster machine.

Table 15. Model comparison of the overall cohort (sub-sample proportion 0.8).

	$AdjustedR^2$
CpG 93 (reference set)	0.9746328
CPG 98 (partition by rows)	0.9756825
CPG 94 (partition by columns)	0.9760258
CPG 93 (partition by blocks)	0.9746377

Figure 28 shows the how the predicted age fit with the true chronical age using different models trained from the CpGs sets selected above. The plot shows the goodness of fit of these models and confirms the efficacy of the core CpGs selection procedure by combining partition methods and sub-sample elastic regression.

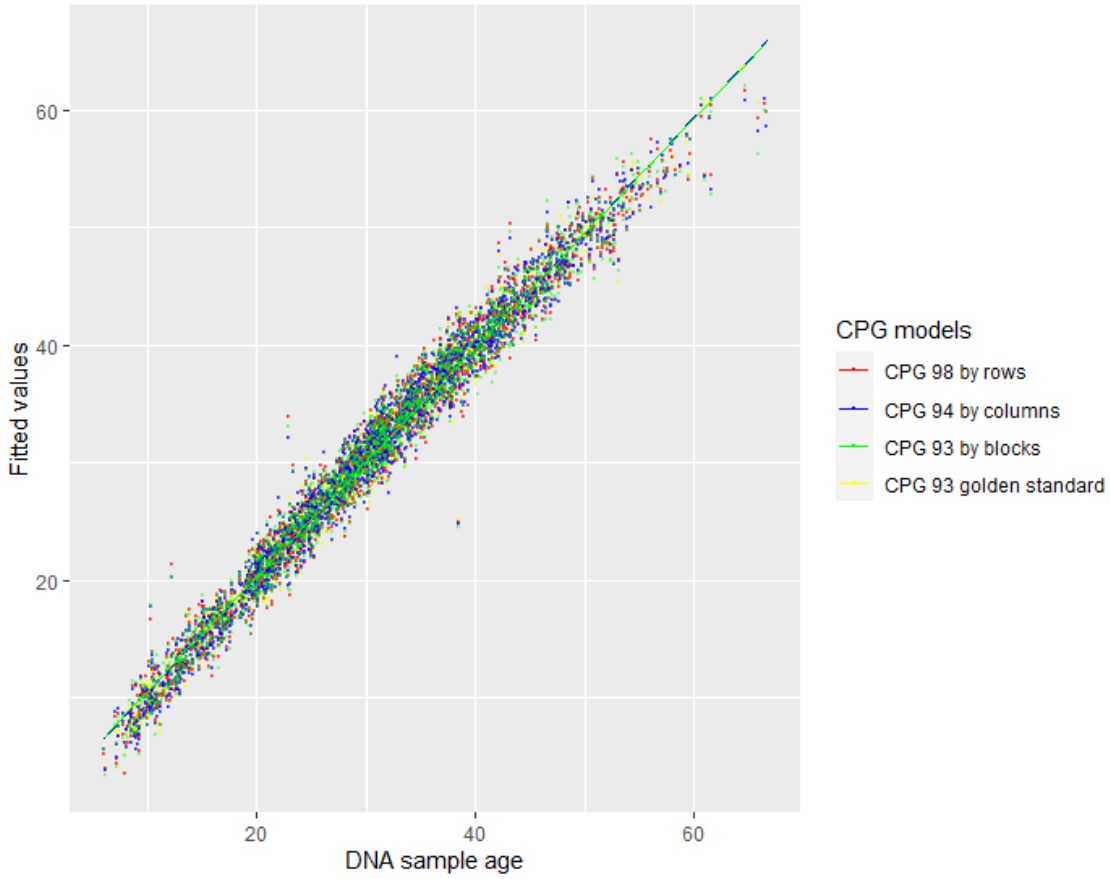

Figure 28. Fitted values vs. DNA sample age using different partition elastic net methods.

Sub-sample proportion 0.5

i. By rows:

The sub-sample elastic net regression is applied to the pre-selected 5,237 CpGs pool created from row partition method. And 32 CpGs are selected (100% recurrence rate) in the final list. Comparing these CpGs with the reference set (29 CpGs), there are 23 CpGs are in common (79.31% of the 29 CpGs). Figure 29 shows the overlap between the 29 CpGs (reference) and the 32 CpGs (partition EN by rows).

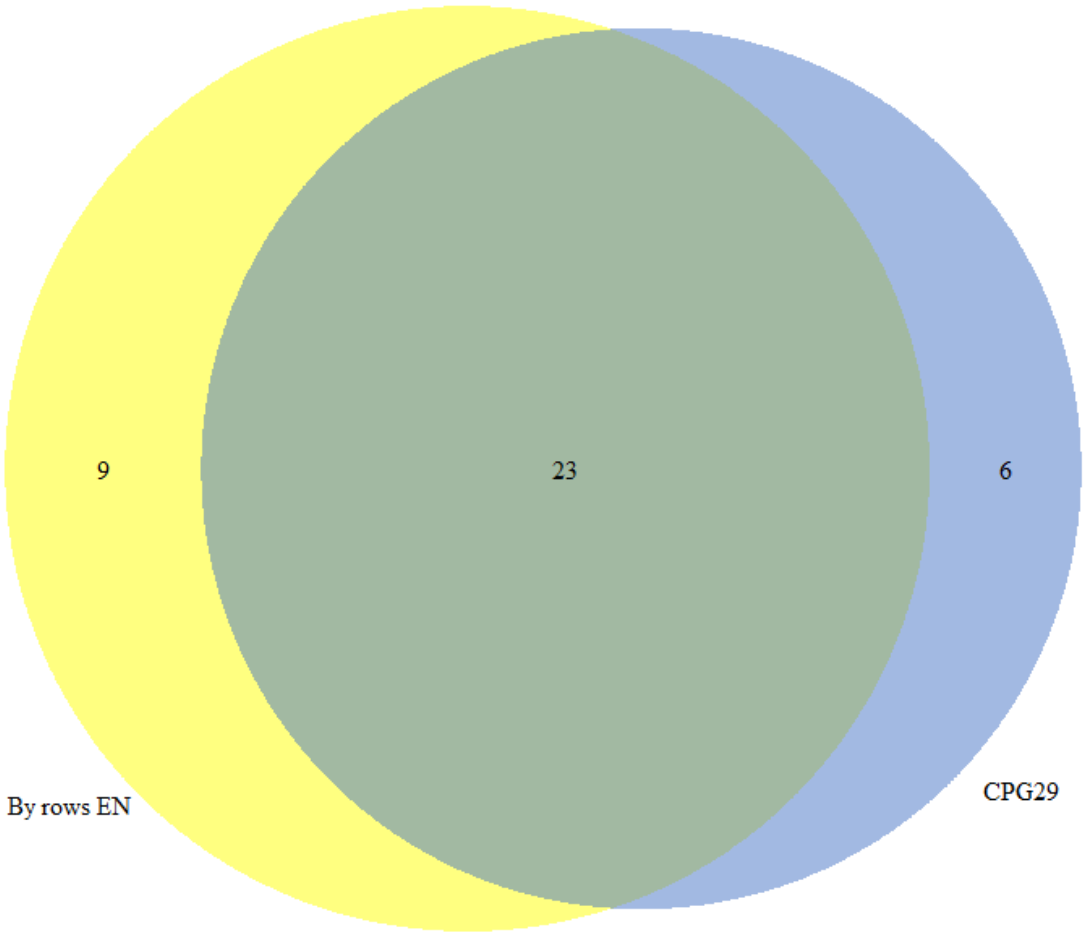

Figure 29. Comparison for the selected 32 CpGs with the reference set (29 CpGs).

ii. By columns:

Based on pre-selected 4,012 CpGs pool, 35 CpGs are selected (100% recurrence rate) after the 10 iterations sub-sample analyses. Comparing the 35 CpGs with the reference CpGs (29 CpGs), there are 21 CpGs are in common (72.41% of the 29 CpGs). Figure 30 shows the overlap between the 29 CpGs (reference) and the 35 CpGs (partition EN by columns).

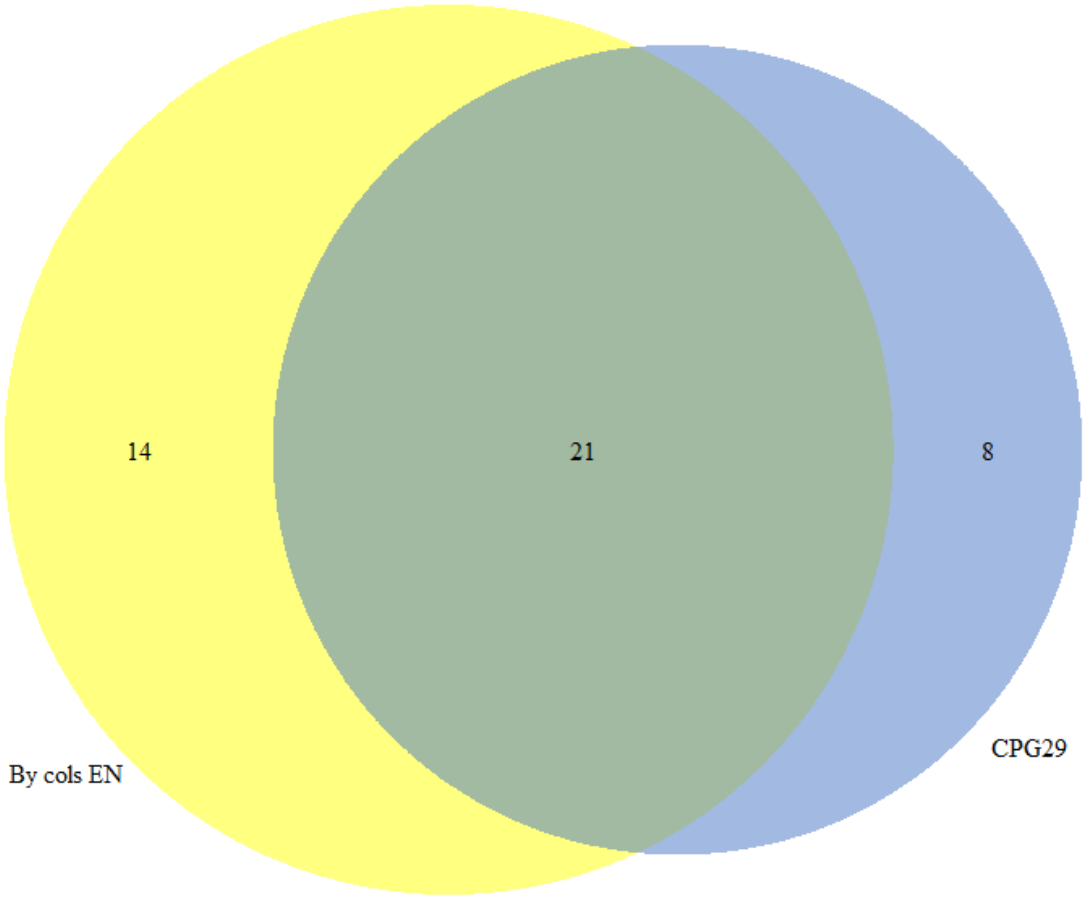

Figure 30. Comparison for the selected 35 CpGs with the reference set (29 CpGs).

iii. By blocks:

By combing the block partition method and sub-sample elastic net regression, 31 CpGs are selected (100% recurrence rate). Comparing these 31 CpGs with the reference (29 CpGs), there are 22 CpGs are in common (75.86% of the 29 CpGs). Figure 31 shows the overlap between the 29 CpGs (reference) and the 31 CpGs (partition EN by blocks).

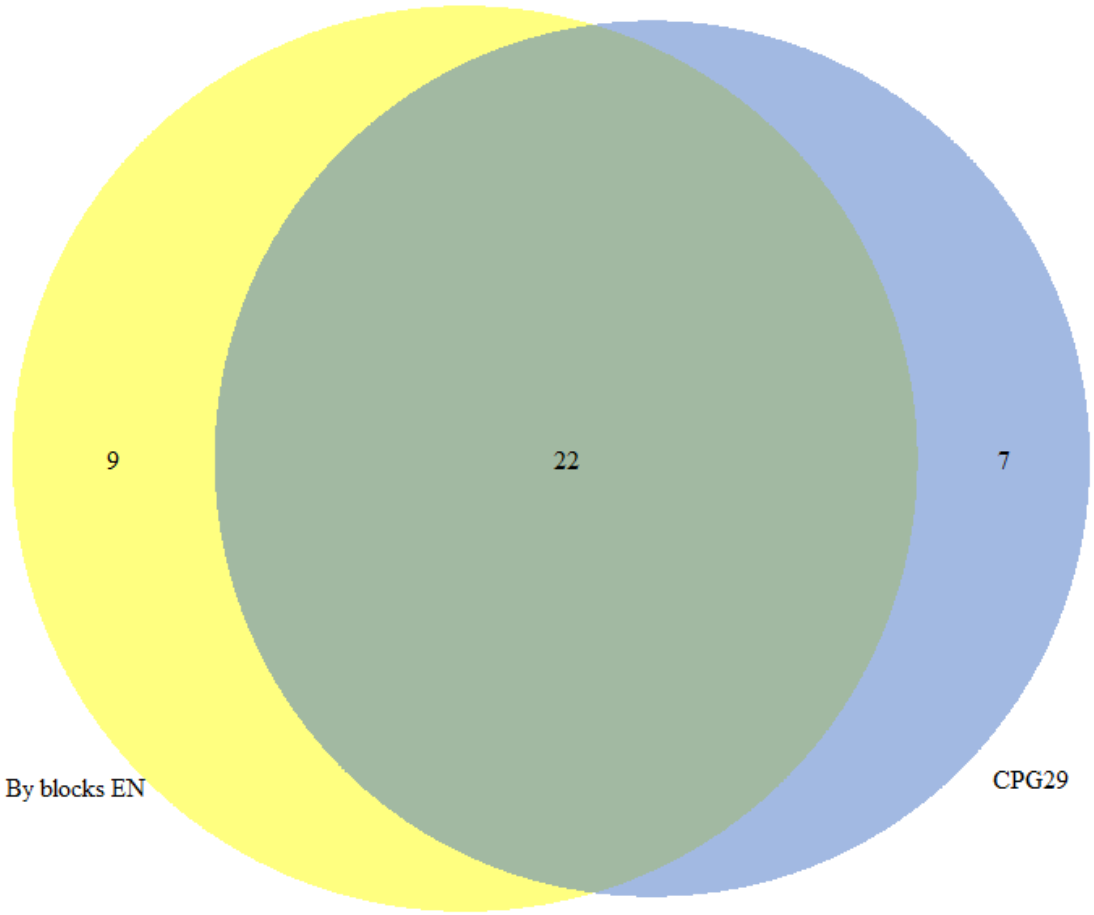

Figure 31. Comparison for the selected 31 CpGs with the reference set (29 CpGs).

When we decrease the sub-sample proportion to 0.5, the performance of core CpGs selection is similar to previous analysis. The results are stable (21 to 23 cores selected) and the row partition method perform best. The Venn-diagram in Figure 32 shows that there are 18 CpGs (62.07% of the 29 CpGs) are in common in the reference 29 CpGs and the other 3 partition methods combing with sub-sample elastic net regression with sample proportion 0.5.

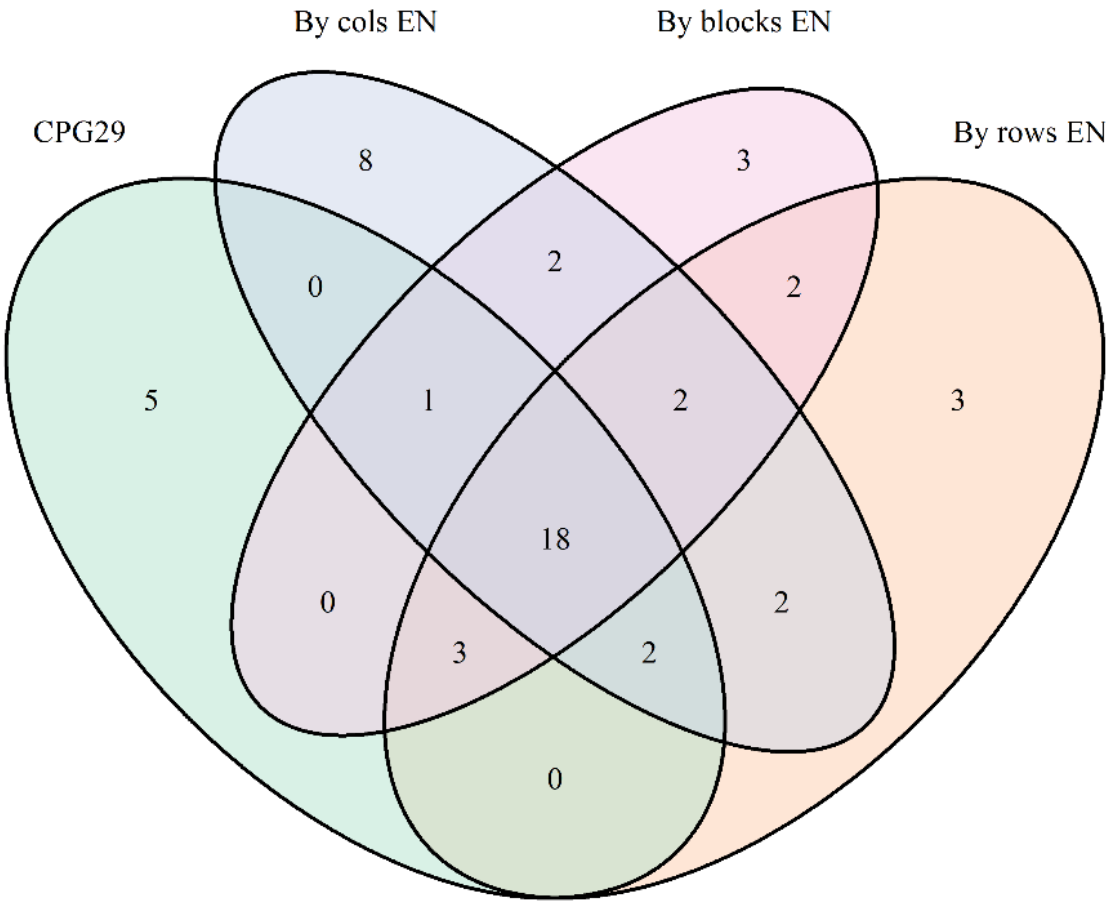

Figure 32. Relationship of the reference 29 CpGs and the CpGs from 3 partition methods with sub-sample proportion 0.5.

Table 16 shows the adjusted R^2 of the trained models using the selected CpGs from reference set and combining the partition methods and sub-sample elastic net methods with proportion 0.5. All of their adjusted R^2's are greater than 0.95, which indicate a goodness of fit of the models with true data.

Table 16. Model comparison of the overall cohort (sub-sample proportion 0.5).

	$AdjustedR^2$
CpG 29 (reference set)	0.9578276
CPG 32 (partition by rows)	0.9582548
CPG 35 (partition by columns)	0.9563559
CPG 31 (partition by blocks)	0.9578049

Figure 33 shows the goodness of fit of the predict age in the trained models with true chronical age.

Figure 33. Fitted values vs. DNA sample age using different partition elastic net methods.

Sub-sample proportion 0.2

i. By rows:

By applying the sub-sample elastic net regression (sample proportion 0.5) to the pre-selected 5,237 CpGs, 10 CpGs are selected (100% recurrence rate). Comparing these CpGs with the reference set (7 CpGs), all the 7 reference CpGs are picked out (100%). Figure 34 shows the overlap between the 7 CpGs (reference) and the 10 CpGs (partition EN by rows).

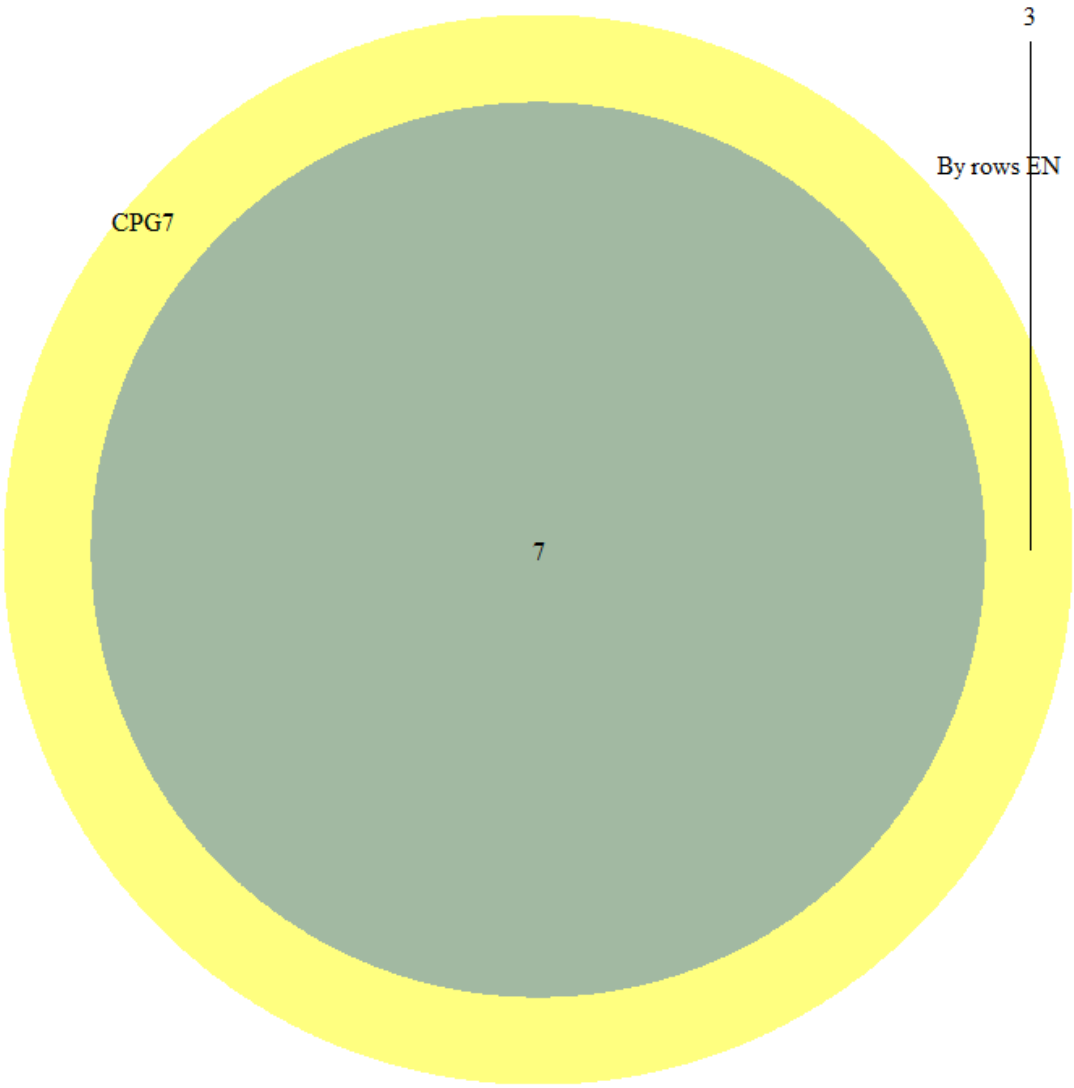

Figure 34. Comparison for the selected 10 CpGs (partitions by rows) with the reference set (7 CpGs).

ii. By columns:

9 CpGs are selected (100% recurrence rate) by combining the column partition method and sub-sample elastic net regression. Comparing the 9 CpGs with the reference set (7 CpGs), there are 5 CpGs that are in common (71.43% of the 7 CpGs). Figure 35 shows the overlap between the 7 CpGs (reference) and the 9 CpGs (partition EN by columns).

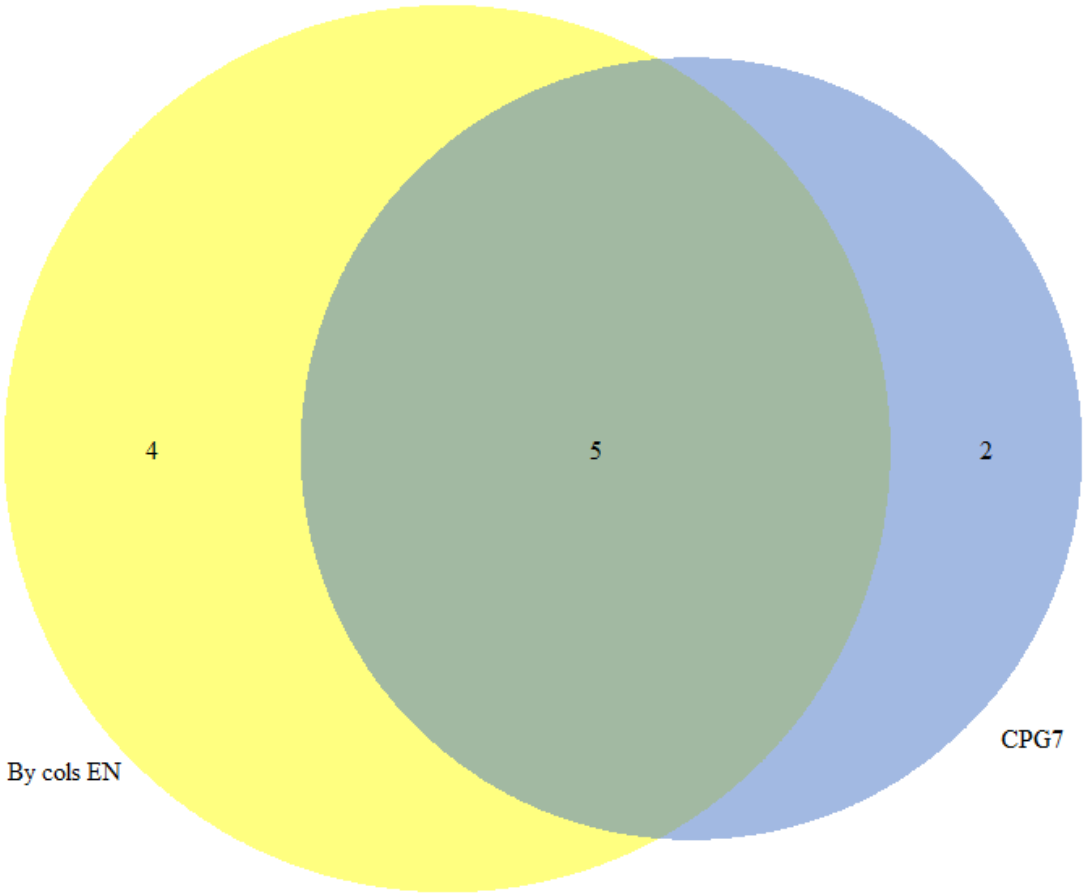

Figure 35. Comparison for the selected 9 CpGs (partitions by columns) with the reference set (7 CpGs).

iii. By blocks:

By combing the block partition method and sub-sample elastic net regression, 7 CpGs are selected (100% recurrence rate). Comparing these 7 CpGs with the reference set (7 CpGs), there are 5 CpGs that are in common (71.43% of the 7 CpGs). Figure 36 shows the overlap between the 7 CpGs (reference) and the 7 CpGs (partition EN by blocks).

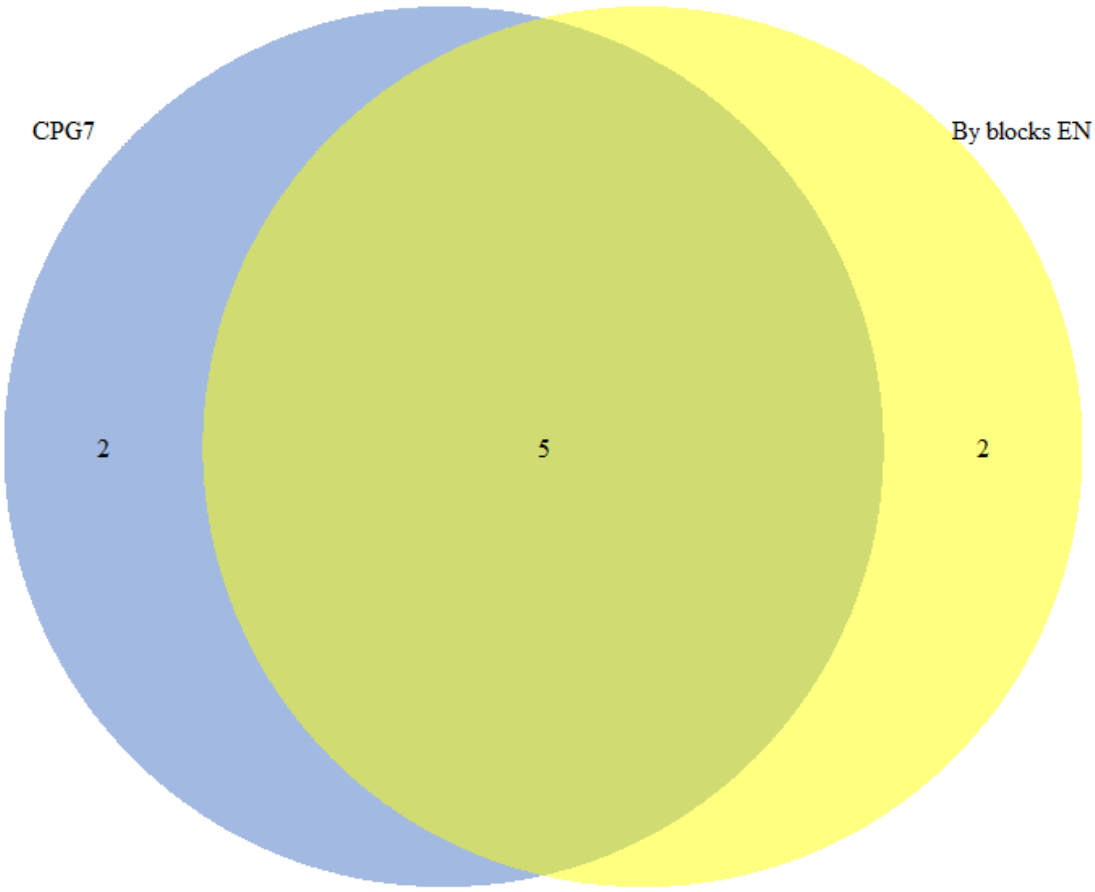

Figure 36. Comparison for the selected 7 CpGs (partitions by blocks) with the reference set (7 CpGs).

When we decrease the sub-sample proportion to 0.2, all of the 3 partition methods can

select most of the core CpGs and the row partition perform best which can screen out all the 7 reference CpGs. The Venn-diagram in Figure 37 shows that there are 4 CpGs (57.14% of the 7 CpGs) are in common in the 7 reference CpGs and the other 3 partition methods combing with sub-sample elastic net regression with sample proportion 0.2.

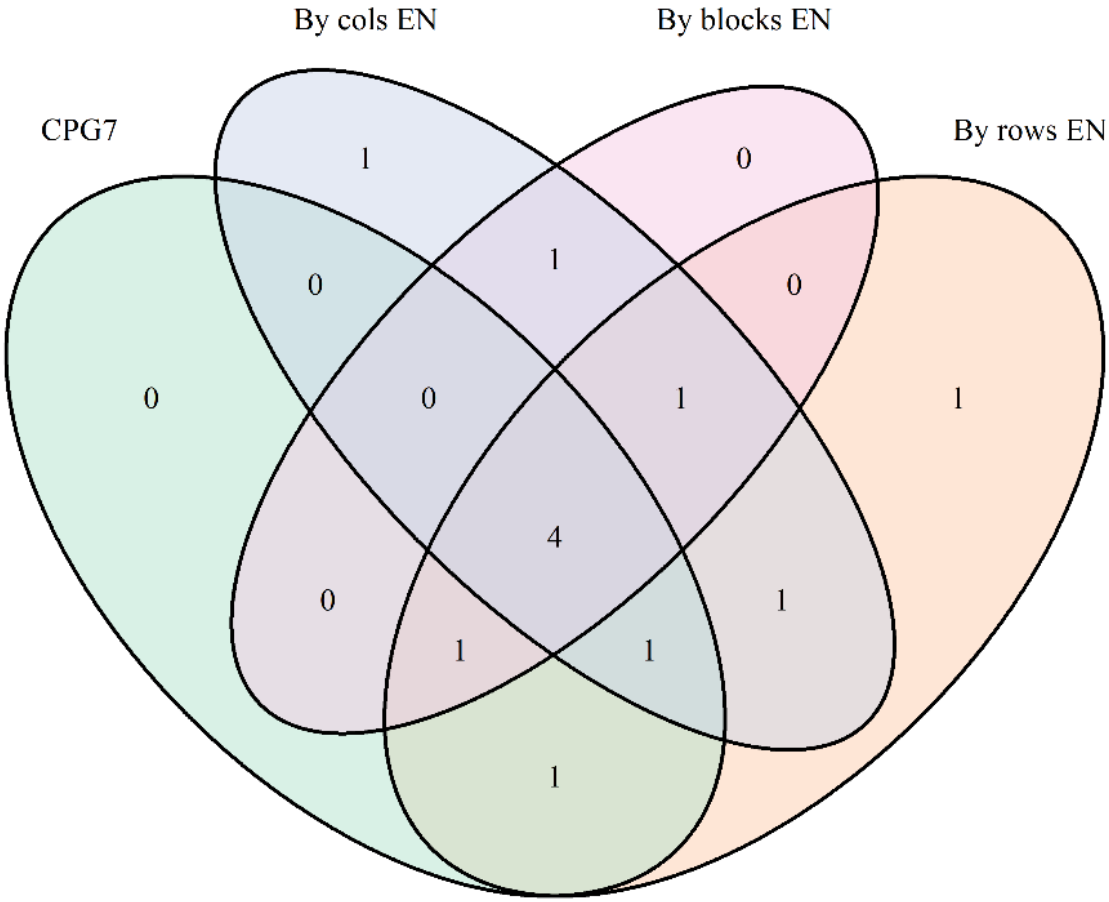

Figure 37. Relationship of the 7 reference CpGs and the CpGs from 3 partition methods with sub-sample proportion 0.2.

Table 17 shows the adjusted R^2 of the trained models using the selected CpGs from refer-

ence set and combining the partition methods and sub-sample elastic net methods with proportion 0.2. All of their adjusted R^2's are greater than 0.89 which indicate the trained models performs well with true data.

Table 17. Model comparison of the overall cohort (sub-sample proportion 0.2).

	$AdjustedR^2$
CpG 7 (reference set)	0.8924991
CPG 10 (partition by rows)	0.9213614
CPG 9 (partition by columns)	0.9174052
CPG 7 (partition by blocks)	0.9069831

Figure 38 shows the goodness of fit of the predict age in the trained models with true chronical age.

Figure 38. Fitted values vs. DNA sample age using different partition elastic net methods.

4.10 Summary of the partition elastic net regression

As showed in Chapter 4, when the memory is not enough to fit an ultrahigh dimension data, the partition elastic method will be a good choice to solve the problem. All of the 3 partition methods perform well and could be an alternative way to select the final variables.

By applying the proposed partition elastic net methods to the CpGs study, the column partition elastic net regression performs very well in overall variables (CpGs) selection. It can screen out 99.50% variables (CpGs) compared to the results when we fit the whole data into the memory. The row partition method and block partition method can screen out 93.69% and 94.85% features (CpGs) respectively comparing to the reference set. And current results are based on 10

iterations, it will be expected with the iteration times increasing the selected variables (CpGs) will be much closer to the reference set (the results when we fit the whole data into the memory).

The combination of partition methods and sub-sample elastic net regression provide an alternative way to select core CpGs when there is a memory issue. By combing with the partition method, the sub-sample elastic net regression can start variable selection from the selected variable pools instead of the original high dimension data. It save computing cost dramatically. Based on the results above, all of the three partitions methods perform stable when we combine with the sub-sample elastic net regression. And the row partition method shows better performance compared to the other two partition methods. All the selected CpGs can build efficient epigenetic clocks with $adj - R^2 s$ range from 0.9070 to 0.9760.

Chapter 5

Other high dimension reduction techniques

Penalized regularized regression are regression based method to select the features which is based on both the outcomes (Y) and predictors (Xs). There are some other dimension reduction methods such as principal component analysis (PCA) which is independent of Y. Utilizing the PCA or a combination of the PCA with linear model may significantly reduce the data dimension. In this Chapter, we apply the PCA to the real high-dimension CpGs data and show it's performance. In addition, another Y independent machine learning method – Autoencoder (AE) is compared to PCA to show their similarity and difference.

5.1 Principal component analysis (PCA)

The regular principal component analysis is a high dimension data reduction technique which mathematically transforms (or defined as orthogonal linear transformation) the large number of variables (p dimension) into equal number of uncorrelated principal components (p). Each component is a linear combination of all the variables in the data. Usually the greatest variance of the scaled data projection lies to the first principal component, the second greatest variance on second principal component, and so on [20]. Thus, the original data variation can mainly be explained by first few numbers of principal components, while the rest can be ignored. The first few major principal components can be used for building predictive models.

To explain the PCA mathematically, we can express the i^{th} principal component as below:

$$y_i = \sum_{j=1}^{p} w_{ij} x_j = W_{(i)}^T \tag{5.1}$$

where w_i is weight vector for i^{th} PC. To maximize the variance of first PC, it requires the condition shown below:

$$W_1 = \text{argmax} \frac{W^T X^T X W}{W^T W} \tag{5.2}$$

where W is the corresponding vector.

For other components, the weight vector can be generally presented as

$$W_k = \text{argmax} \frac{W^T \tilde{X}_k^T \tilde{X}_k W}{W^T W} \tag{5.3}$$

for the k^{th} PC, where

$$\tilde{X}_k = X - \sum_{i=1}^{k-1} X W_{(i)} W_{(i)}^T \tag{5.4}$$

The transformed data $PCs = XW$ are uncorrelated, and we can reconstruct X by

$$X = PCs \times W^T \tag{5.5}$$

If we use the first k PCs to reconstruct the X, we can get a rough estimate of original X as

$$\hat{X} = PCs_{(1\,to\,k)} \times W_k^T \tag{5.6}$$

5.2 PCA application in real high dimension data (CpGs data)

Different from penalize regression, principal component analysis is used for dimension reduction but not select the variables. When facing an ultra-high dimension data, should we directly use the PCA method or combine it with penalized regression needs to be studied.

i. Apply PCA on the overall CpGs data (689,414 CpGs)

Firstly, a regular principal component analysis is applied to the overall CpGs data which include 2,112 patients and 689,414 CpGs. There are 689,414 principal components created. Each PC may only explain a small portion of the data variation. The figure 39 shows the data variation explained by each PC and their cumulative variation proportion.

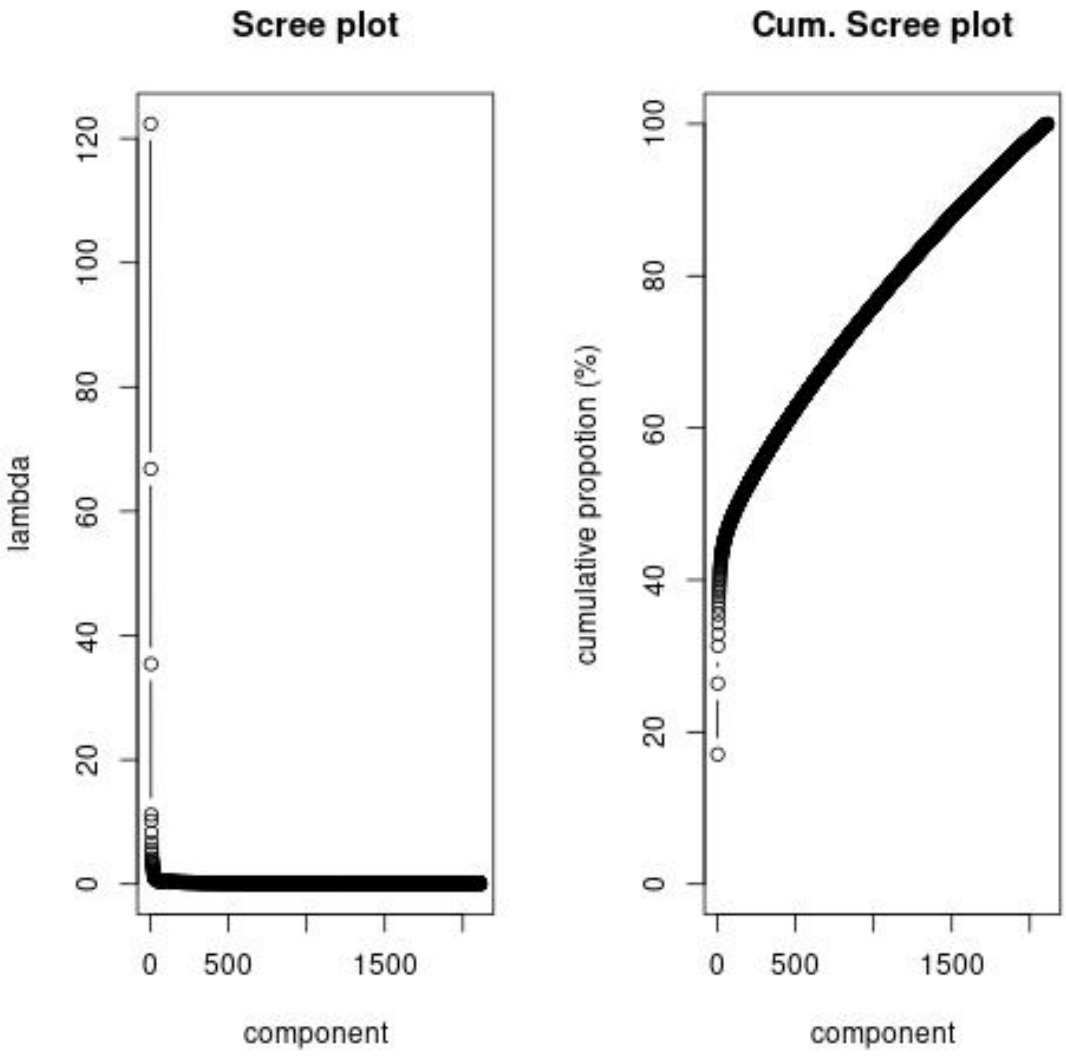

Figure 39. Data variation explained by principal components analysis (overall PCA).

Due to the huge dimension of the PCA matrix 689,414×689,414, we only keep the first 200 rank of principal components. And we tried to use the first few PCs to build the model to predict

the chronical age. However, The first 200 PCs can only present 52.66% of the data variation. Thus, only using the first few numbers of PCs to build the model may lose data information and have a model fit problem. Table 18 shows the number of principal components and the adjusted R^2 of trained models using these PCs. The results show us that we need 50 PCs to get an acceptable model for chronical age prediction.

Table 18. Adjusted R^2 of trained models using different numbers of PCs (overall PCA).

	$AdjustedR^2$
First 1 PC	0.2013
First 2 PCs	0.3867
First 3 PCs	0.3952
First 5 PCs	0.6130
First 10 PCs	0.8338
First 50 PCs	0.9255

ii. Combine the PCA with elastic net regression

Although the principal component analysis shows a significant dimension reduction in CpGs data analysis (from 689,414 to 50), statistically it still looks high comparing to the sample size is only 2,112. We explored the possibility to further reduce the dimension by combining the PCA technique with elastic net regularized regression.

As we showed in chapter 2, applying the elastic net regression on the overall data selected 602 CpGs. The principal component analysis is applied to the selected 602 CpGs data (2,112 patients). The figure 40 shows the data variation explained by each PC and their cumulative variation proportion.

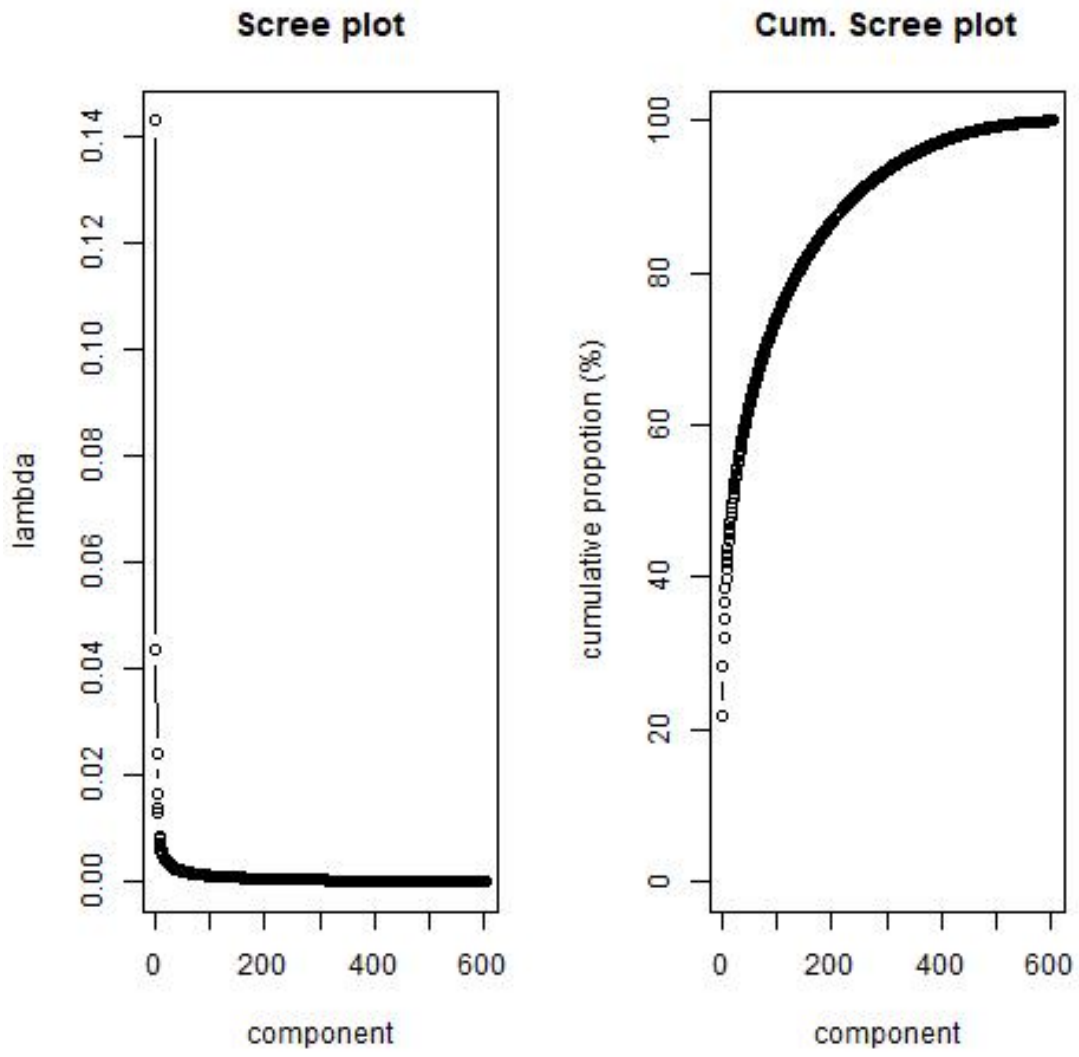

Figure 40. Data variation explained by principal components analysis (PCA for 602 CpGs data).

Similar to the results of overall PCA, 20 PCs are needed to explain the 50% of the data variation. However, the 602 CpGs are pre-screened and are more highly related with the outcome "chronical age". It is expected to have a better model fit when we use the PCs from the 602 CpGs. Table 19 shows the number of principal components and the adjusted R^2 of trained models using these PCs. The results showed us that we only need the first PC to get a good model (adj R^2=0.9467) for chronical age prediction and first 10 PCs to get a better model fit the CpGs data

very well (adj R^2=0.9685). It indicates that the combination of PCA and elastic net regression will be a potential way to deal with high dimension data in building the prediction models and data analysis.

Table 19. Adjusted R^2 of trained models using different numbers of PCs (PCA for 602 CpGs data).

	$AdjustedR^2$
First 1 PC	0.9467
First 2 PCs	0.9467
First 3 PCs	0.9469
First 5 PCs	0.9664
First 10 PCs	0.9685
First 50 PCs	0.9808

5.3 Autoencoder vs PCA

An autoencoder belongs to the artificial neural network family [21]. Typically, it consists of 3 parts: "Input layer", "bottleneck layer (code layer)" and "output layer". And there are two major steps: "encode" which maps the input layer to code layer and "decode" which maps the code layer to output layer (figure 41 shows the autoencoder structure).

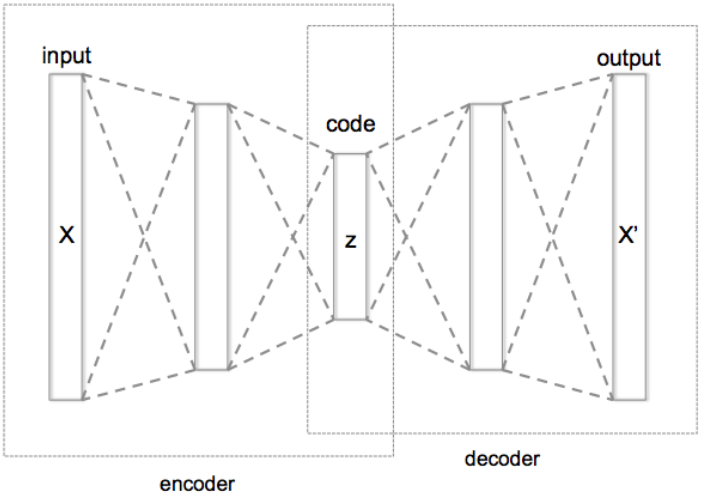

Figure 41. Autoencoder structure (https://en.wikipedia.org/wiki/Autoencoder).

Similar to PCA, the autoencoder is only related with independent variables but not response variables. It reconstructs the input X into X' with the same dimension and tries to minimize the reconstruct error. Though, the PCA and autoencoder looks similar, there are a few major differences summarized in table 20.

Table 20. Major difference between PCA and AE.

PCA	AE
Handles linear transformation	Can handle non-linear functions
PCs are linearly uncorrelated	AE features might be correlated
Faster and computationally cheaper	May be slower and require more sources
More generalized	May have over-fitting problem

It is widely believed that a single layered autoencoder with linear activation function is very similar to PCA. It should not be exactly the same, as the AE is trained for accurate reconstruction and the features may be correlated while the PCA has uncorrelated PCs. To illustrate the similarity and the slightly difference between PCA and AE, we used the 7 CpGs data to compare the PCA and AE techniques. We construct the linear autoencoder by using "adam" optimization and "squared"

loss type with one single layer (using linear activation function, R package ANN2). The PCA reconstruction is based on the formula:$\hat{X} = PCs_k \times W_k^T$, where PCs_k is the principal component matrix keep only the first k principal components, W_k^T is the transpose matrix of the first k eigenvectors. In figure 42 the reconstruction errors (MSE) were compared between linear autoencoder and PCA for different numbers of PCs/code layer Nodes (1 to 7). The results showed us both the linear autoencoder and the PCA have very small MSE which indicate pretty good reconstruction accuracy. However, the PCA performs better in this example and there exists difference between PCA and linear autoencoder.

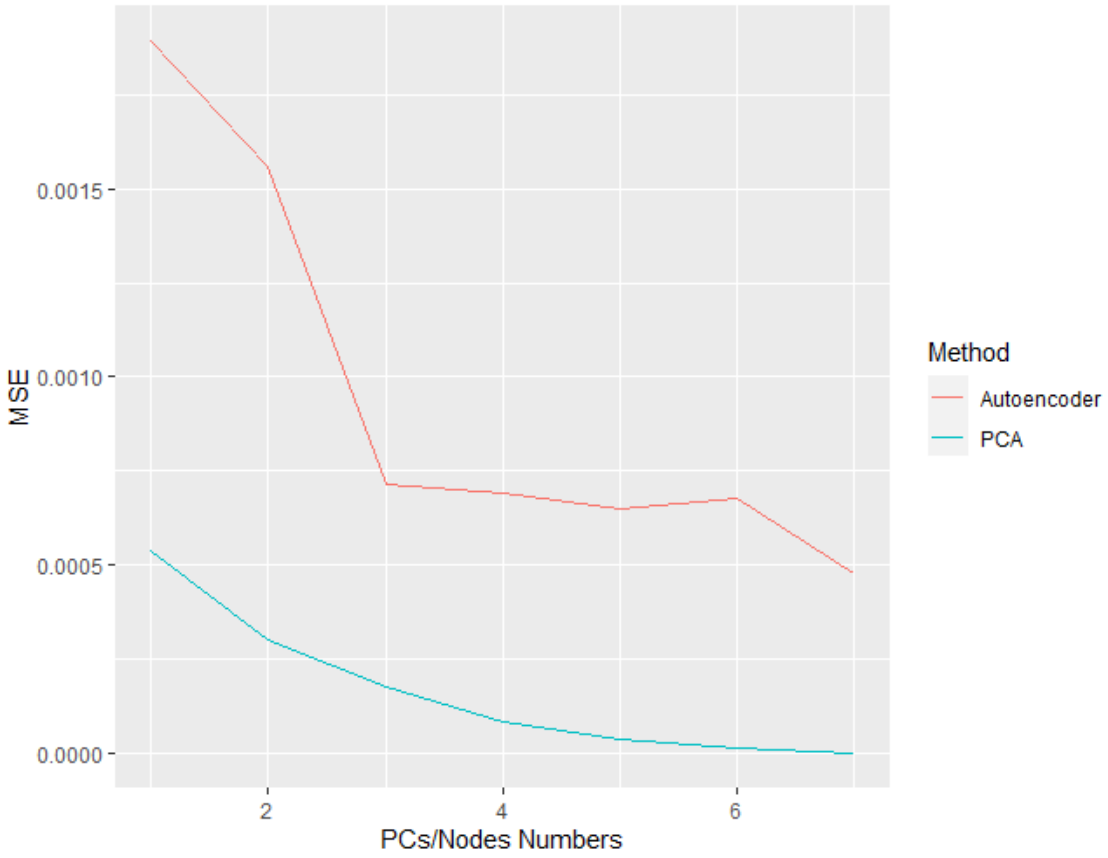

Figure 42. Comparison of reconstruction errors between PCA and AE.

Chapter 6

Sensitivity analysis of different training cohort in variables selection

In a well designed study, the training cohort and validation cohorts should be balanced and share the similar characteristic, which will lead to robust result in analysis. However, many unknown factors are associated with training and validation data. A random split training cohort cannot account for these factors and will leads to unstable results.

To show the variation in variable selection according to different training data, the sensitivity analysis is performed on the CpGs study.

6.1 Variation in CpGs selection among different training cohort

In the study of CpGs data, the CpGs are highly correlated and easy to find replaceable CpGs when the training data varies. As showed in Chapter 2, different training cohort will select different CpGs (602, 589, 604, 595, 611 and 598 respectively). Among the 6 selected CpG sets, 141 CpGs were repeatedly selected (figure 1). To evaluate the robustness of the initially selected core CpGs, they were compared to the other 5 selected CpGs set (excluding the 602 CpGs) in the following figures.

Figure 43 shows that among the 93 CpGs, 64 CpGs were repeatedly selected by the other 5 CpGs sets.

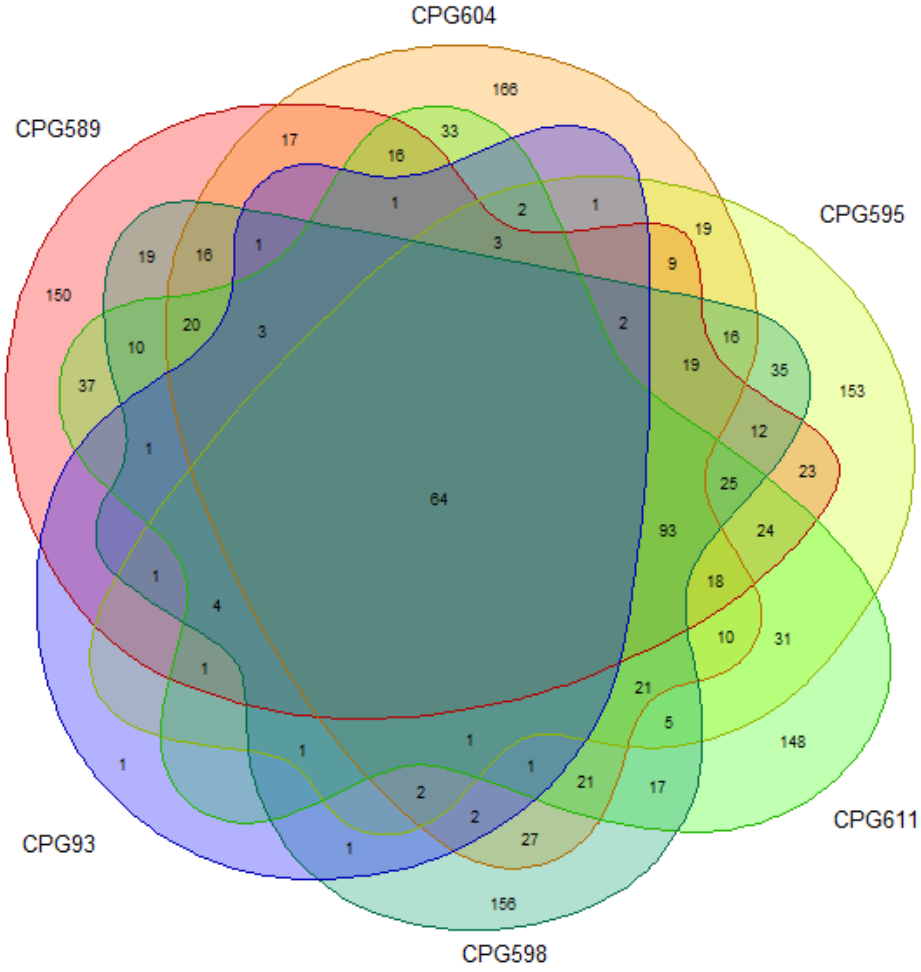

Figure 43. CpG 93 vs other 5 CpG sets

When comparing the core 29 CpGs and 7 CpGs with the other 5 CpG sets, there were 28 and 7 CpGs in common respectively (figure 44 and figure 45).

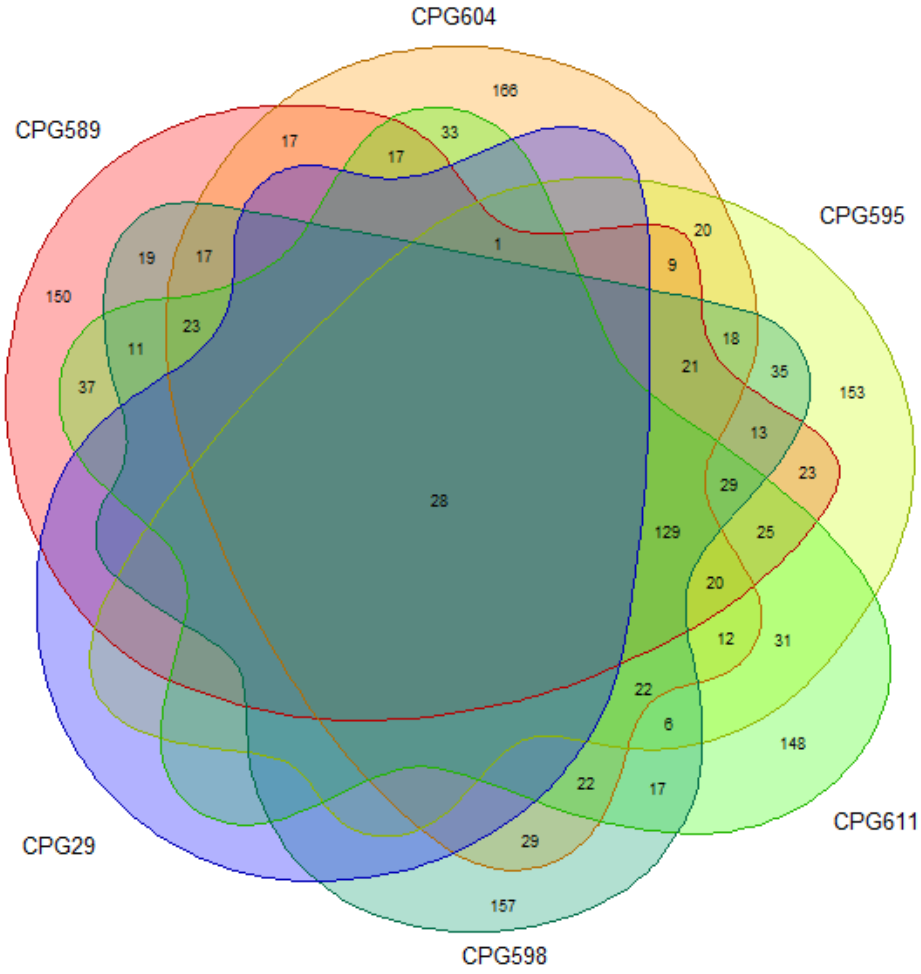

Figure 44. CpG 29 vs other 5 CpG sets

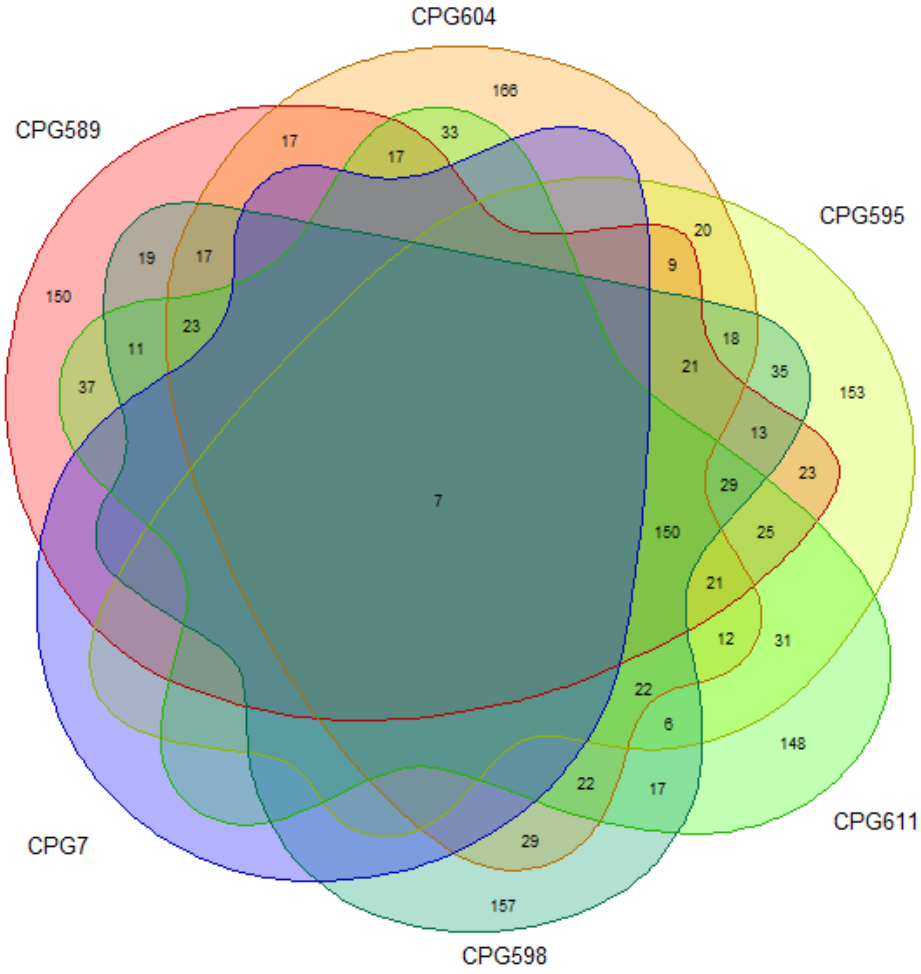

Figure 45. CpG 7 vs other 5 CpG sets

These results confirm that the core CpGs are repeated selected by a different split training cohort. And using the 93, 29 7 CpGs aging clock will provide a relative stable performance in chronical age prediction.

6.2 Core CpGs variation among different training cohort

To further confirms the robustness of the core CpGs selection in the differently split training cohort, we select the 589 CpG model's training cohort ($2,112 \times 689,414$) to do a sensitivity analysis using sub-sampling elastic net regression method (100% recurrence rate as selection criteria). We select this cohort because it's performance ranks at the bottom among the 6 models shown in table 2 and may have the maximum deviant from the results based on the 602 CpG model's training cohort.

Sample proportion 0.8

97 CpGs were repeatedly selected (100% recurrence rate) from the 689,414 CpGs based on the new training data. And 58 (62.37%) of them overlap with the originally selected core 93 CpGs (showed in figure 46). The developed model with 97 CpGs has $ajd - R^2$ 0.9770.

Figure 46. CpGs 97 vs reference CpGs 93

Sample proportion 0.5

30 CpGs were repeatedly selected (100% recurrence rate) from the 689,414 CpGs based on the new training data. And 15 (51.72%) of them overlap with the original selected core 29 CpGs (showed in figure 47). The developed model with 30 CpGs has $ajd - R^2$ 0.9537.

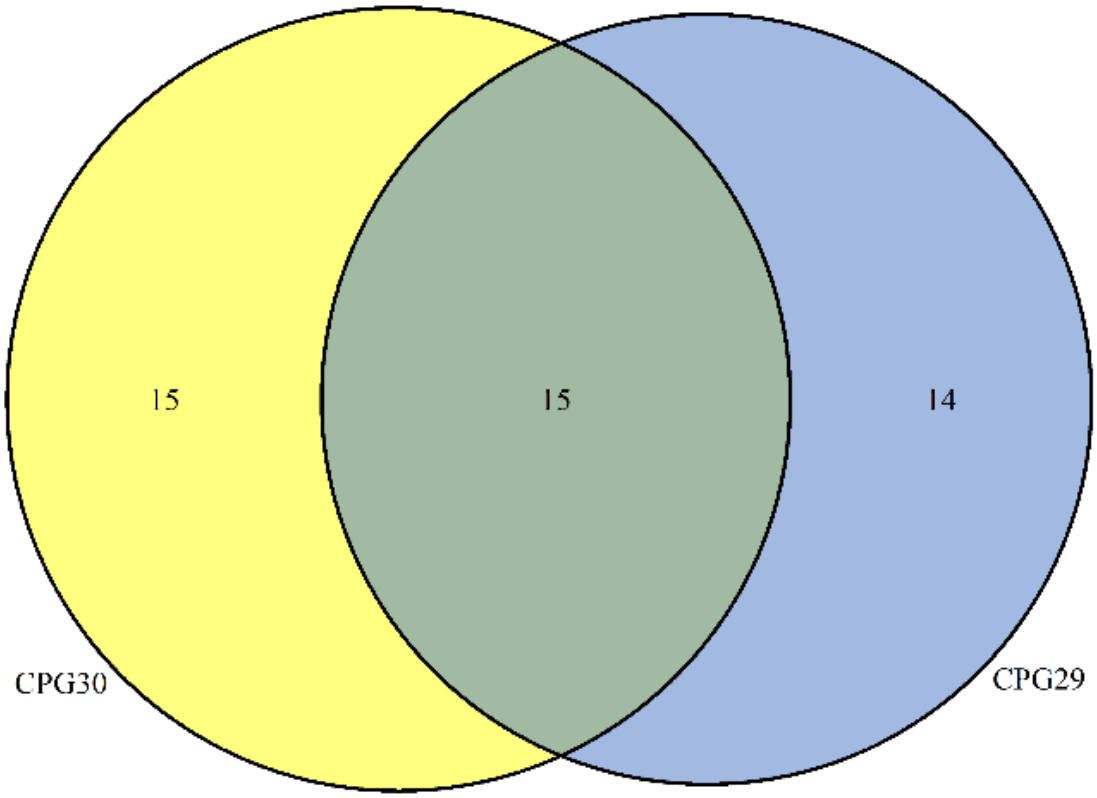

Figure 47. CpGs 30 vs reference CpGs 29

Sample proportion 0.2

7 CpGs were repeatedly selected (100% recurrence rate) from the 689,414 CpGs based on the new training data. And 6 (85.71%) of them overlap with the originally selected core 7 CpGs (shown in figure 48). The developed model with 7 CpGs has $ajd - R^2$ 0.8881.

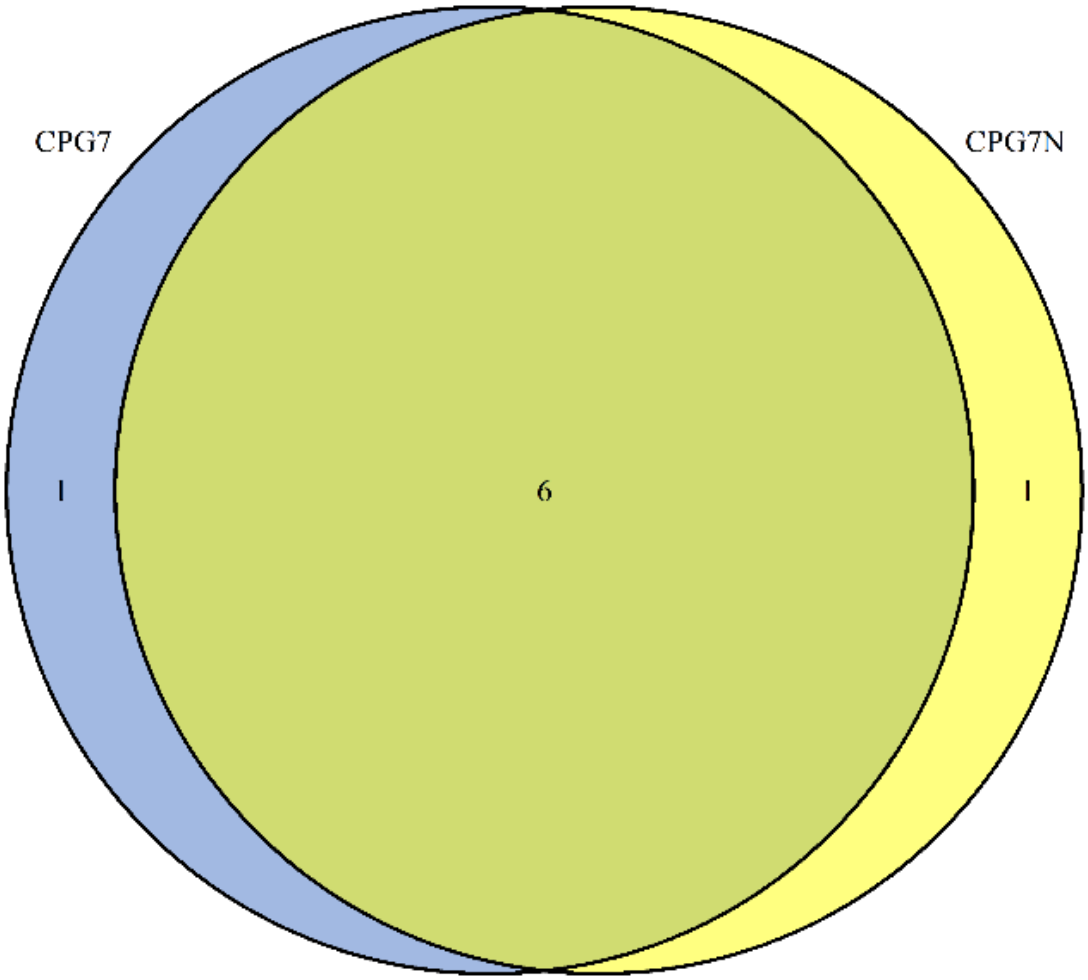

Figure 48. CpGs 7 vs reference CpGs 7

The venn-diagram below(figure 49) shows the relationship of the newly selected core CpG sets with the 589 CpG set. It has a similar pattern to what we showed in chapter 2 which has all the CpG sets connected to each other.

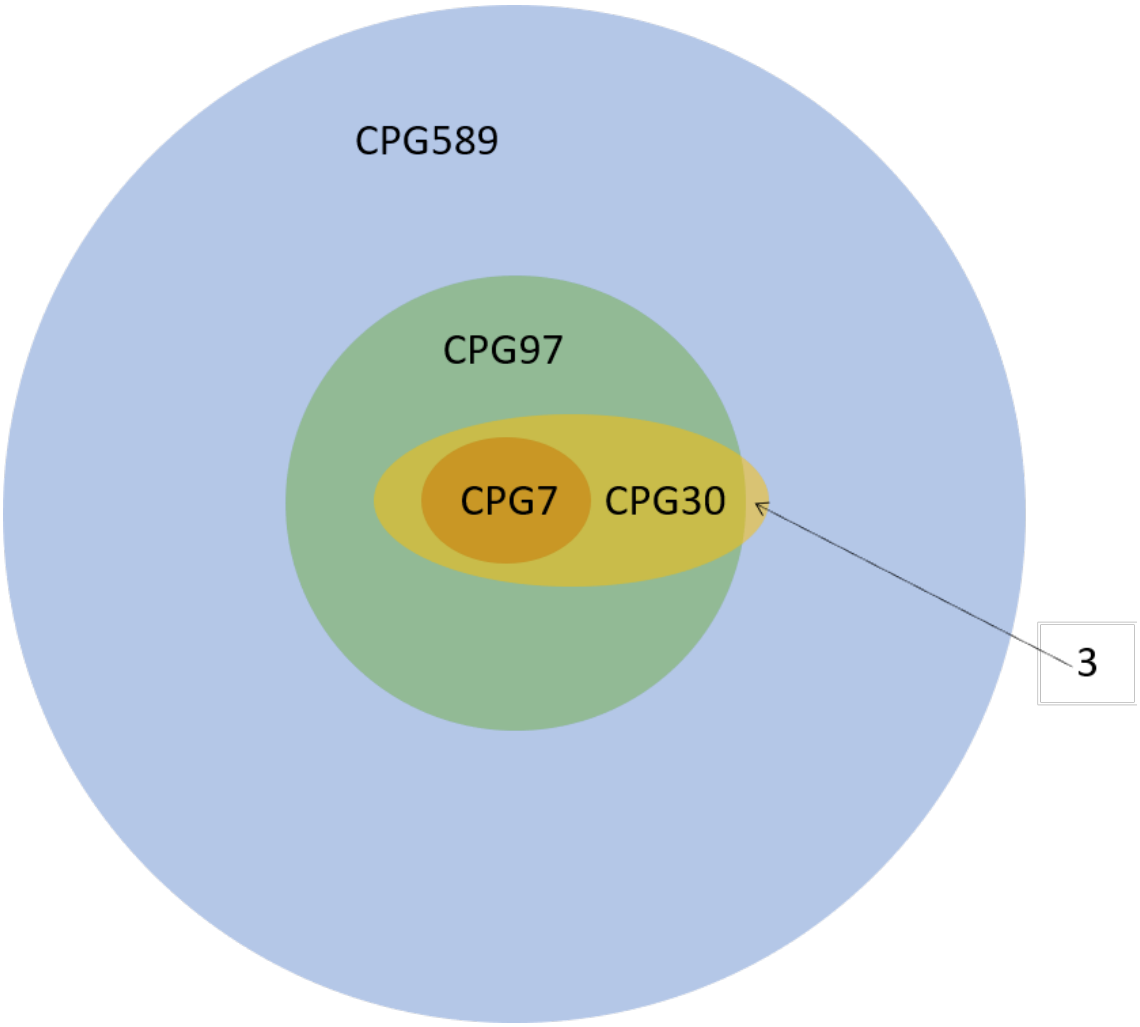

Figure 49. Relationship of the selected CPG sets from the new training cohort

6.3 Epigenetic colocks performance comparison

To evaluate the CpGs models developed from the new training cohort, the newly trained 4 models are applied to SJLIFE validation and controls cohort. A comparison of the newly developed models performance with original results are showed below:

Table 21. Validate the models in validation and control data.

	Validation (Survivor)				Control				Survivors vs Controls	Survivors vs Controls
	MSE	RMSE	Beta (slope)	Intercept	MSE	RMSE	Beta (slope)	Intercept	Beta comparison P-value	Intercept comparison P-value
CpG 602	4.16	2.04	0.972	0.878	4.27	2.07	0.922	2.524	0.359	0.022
CpG 93	4.66	2.16	0.973	0.828	4.87	2.21	0.916	2.859	0.344	0.008
CpG 29	6.93	2.63	0.982	0.464	6.12	2.47	0.902	3.232	0.446	$<$0.0001
CpG 7	14.21	3.77	0.917	2.513	13.85	3.72	0.808	7.352	0.272	$<$0.0001
CpG 589	4.16	2.04	0.959	1.329	4.94	2.22	0.912	2.999	0.371	0.0299
CpG 97	4.69	2.17	0.951	1.591	5.22	2.29	0.916	3.118	0.404	0.0371
CpG 30	7.46	2.73	0.941	2.352	6.64	2.58	0.904	3.550	0.451	0.0973
CpG 7	14.13	3.76	0.877	3.898	13.98	3.74	0.818	6.672	0.372	0.0050

The two sets of models perform similar while the original one is slightly better with smaller RMSE as well as their beta coefficients are closer to 1. Figure 50 visualizes the models performance of the newly developed chronical age clocks in validation and control cohorts which have pretty good fitness.

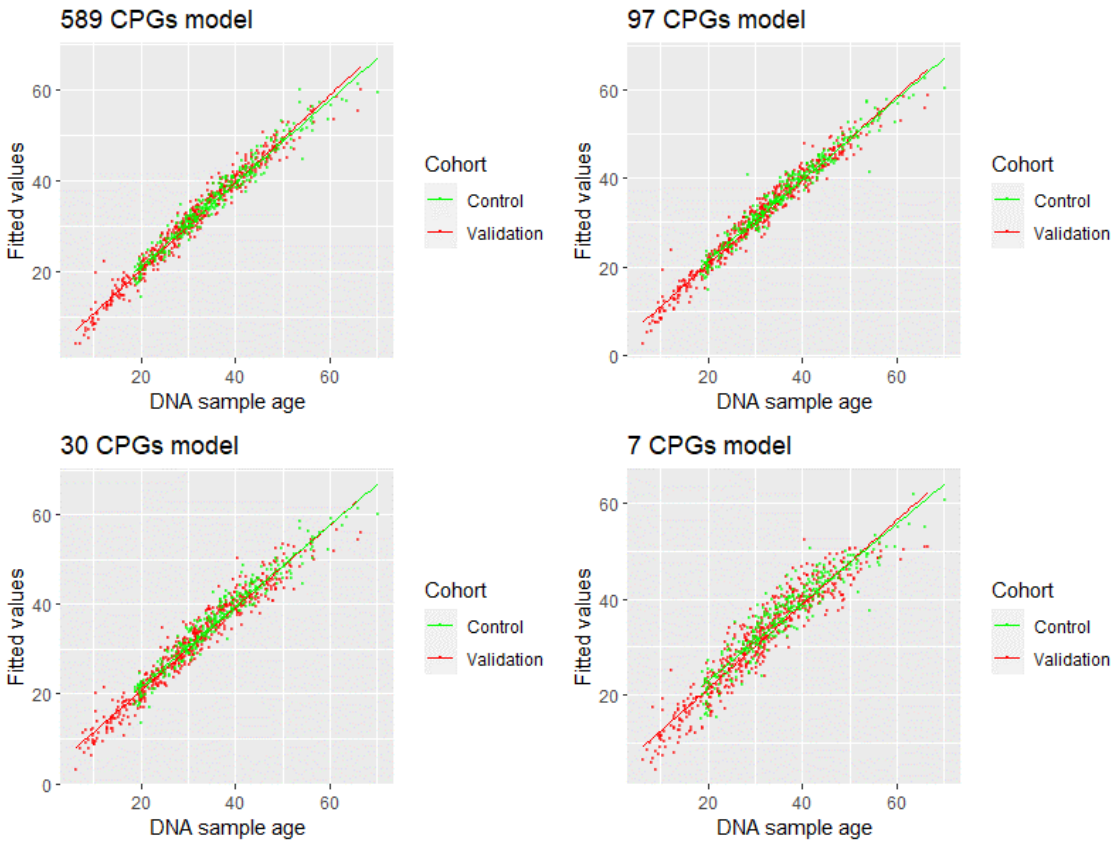

Figure 50. Models validation in validation and control data

6.4 Summary of the impact of different training cohort in CpGs selection

In this chapter, it is noticed that there is CpGs selection variation based on different training cohort. Some of the originally selected CpGs are repeatedly selected basing on different training cohorts especially for the core CpGs. It means that the core CpGs are truly important to build the epigenetic colock. The sensitivity analysis is performed on the newly split training cohort in core CpGs selection. The selected core CpGs are 51.72% to 85.71% overlap with the originally selected core CpGs for different sub-sampling proportions.

Though, the CpGs selection variance exists, the newly developed epigenetic clocks perform similar as the original selected clocks. This is mainly because the CpGs are highly correlated and it is very easy to find replace CpGs while the training cohort changing. The 689,414 CpGs can

build millions of different chronical age clocks and performs well among different training and validation cohort. But it is clear that the core CpGs selection is relatively robust while using sub-sampling elastic net regression method with small sample proportion.

Chapter 7

Conclusion and future study

Elastic net regularized regression is a well established method in feature selection for high-dimension data. An epigenetic aging clock containing 602 CpGs was built by this method in childhood cancer survivors. This clock outperforms the 4 published epigenetic clocks (Hannum, Hovarth, Levine, GrimAge clocks) in chronic age prediction and shows possible generalize using in both cancer survivors and health controls.

The sub-sample elastic net regularized regression provide a way to further reduce the number of features in high-dimension data study. A smaller sample proportion and a higher recurrence rate as inclusion criteria will select the more important features/variables from the training data. In real application, sub-sample elastic net method selected 93, 29, and 7 core CpGs from the original 689,414 features as the sample proportion decrease. The 7 CpGs were repeatedly selected by other sample proportion settings. The trained model from the 7 CpGs has adj-R^2 of 0.89, with prediction deviance of 3.77 years from true value.

The partition elastic net regression methods provide alternative ways to perform elastic net regression when the data is too big to be fitted into the memory. By randomly dividing the data into small sub-datasets, the memory requirement will decrease for each sub-analysis. Multiple iterations of the partition methods can minimize the random partition bias. As shown in real data application, the column partition method can select 99.50% CpGs compare to the CpGs set selected by fitting the whole data into the memory. And the row partition method performs best in the core CpGs selection combining with the sub-sample elastic net regression.

The PCA analysis is independent of the outcome (Y) and performs well in dimension reduction. A combination of the elastic net regression and PCA technique performs better than only performing the PCA. By applying the PCA on the selected CpGs from the elastic net regression analysis, the first principal component can produce an acceptable model with adj-R^2 of 0.9467. It is amazing that the dimension can be reduced from 689,414 to 1 with good model fitness. When

10 principal components included, the adj-R^2 increases to 0.9685 which is very close to a perfect model.

A further comparison of PCA and Autoencoder show both techniques perform well in data reconstruction and dimension reduction. In current data, CpGs show strong linear association with chronical age and performs better than AE. However, AE can deal with non-linear associated data and will perform good non-linear outcomes.

Different training cohort leads to the CpGs selection variation. The core CpGs selection is relative stable with overlap rate range from 51.72% to 85.71% and performs well with similar $adj - R^2$ to the original selected core CpGs. With a small sample proportion the sub-sampling elastic net regression will select stable core CpGs among different training cohort.

In the future study we may study more on how to use the elastic net regression, kernel PCA or other deep learning methods to select features from non-linearly associated data and build efficient prediction models. And the partition methods can be refined and widely used when the memory issue exists, and a potential R package can be developed based on it.

Bibliography

[1] Li Z. Song N. "Shortened Leukocyte Telomere Length Associates with an Increased Prevalence of Chronic Health Conditions among Survivors of Childhood Cancer: A Report from the St. Jude Lifetime Cohort." In: *Clinical Cancer Research* (2020).

[2] Li Z. Qin N. "Epigenetic age acceleration and chronic health conditions among adult survivors of childhood cancer." In: *Journal of the National Cancer Institute* (2021).

[3] Song N. "Persistent variations of blood DNA methylation associated with treatment exposures and risk for cardiometabolic outcomes in long-term survivors of childhood cancer in the St. Jude Lifetime Cohort." In: *Genome Medicine* (2021).

[4] Hannum G. "Genome-wide methylation profiles reveal quantitative views of human aging rates." In: *Mol Cell.* (2013).

[5] Horvath S. "DNA methylation age of human tissues and cell types". In: *Genome Biol.* (2013).

[6] Levine E. "Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning." In: *Aging* (2015).

[7] Lu AT. "DNA methylation GrimAge strongly predicts lifespan and healthspan." In: *Aging* (2019).

[8] *Analyze Illumina Infinium DNA methylation arrays*. `https://bioconductor.org/packages/release/bioc/html/minfi.html`. Bioconductor.

[9] David A. F. *Statistical Models: Theory and Practice.* 2009.

[10] Alvin C. Rencher and G. Bruce Schaalje. *Linear models in statistics.* 2008.

[11] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." In: *Technometrics* (1970).

[12] Zhang H. *Analyzing High-Dimensional Gene Expression and DNA Methylation Data with R.* 2020.

[13] Robert. Tibshirani. “Regression Shrinkage and Selection via the lasso.” In: *Journal of the Royal Statistical Society* (1996).

[14] Hastie Trevor Zou Hui. “Regularization and Variable Selection via the Elastic Net.” In: *Journal of the Royal Statistical Society* (2005).

[15] Akaike H. “A new look at the statistical model identification.” In: *IEEE Transactions on Automatic Control* (1974).

[16] G. Schwarz. “Estimating the dimension of a model.” In: *The Annals of Statistics* (1978).

[17] Anderson D. Burnham K. *Model Selection and Multimodel Inference: A practical information-theoretic approach.* 2002.

[18] Vrieze S. “Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).” In: *Psychological Methods* (2012).

[19] Peterson T. Aho K. Derryberry D. “Model selection for ecologists: the worldviews of AIC and BIC.” In: *Ecology* (2014).

[20] I. T. Jolliffe. *Principal Component Analysis.* 2002.

[21] Mark A. Kramer. “Nonlinear principal component analysis using autoassociative neural networks.” In: *AIChE Journal* (1991).