A multi-tissue age prediction model based on DNA methylation analysis

Hussain Alsaleh[a*], Nicola A. McCallum[a], Daniel L. Halligan[b], Penelope R. Haddrill[a]

[a] *Centre for Forensic Science, Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland, UK*
[b] *Fios Genomics, Nine Edinburgh Bioquarter, 9 Little France Road, Edinburgh, Scotland UK*

[*] Corresponding author. Tel:+44 141 548 5992 ext.: 4519; fax: +44 141 548 2532;
E-mail: hussain-alsaleh@strath.ac.uk

Abstract

Age related tissue-specific DNA methylation markers have been identified in many studies, which can be used to estimate the chronological age of an unknown biological sample's donor. However, if these markers have been used on the wrong type of tissue, they will give an inaccurate age estimation. This research has therefore examined HumanMethylation450 (HM450) BeadChip-based profiles retrieved from the NCBI repository, with the aim of identifying a set of universal DNA methylation markers across forensically relevant tissues. By using elastic net regression, it was possible to identify 10 age-related (AR) DNA methylation markers across 41 samples coming from five types of tissue (whole blood, saliva, semen, menstrual blood, and vaginal secretions). The average predictive accuracy of the constructed model based on training data is 3.8 years. In an independent dataset of 24 samples from four types of tissues (blood, saliva, menstrual blood, and vaginal secretions), the mean absolute deviation for the menstrual blood and vaginal fluid is 6.9 years, 5.6 years for buccal swabs, and 7.8 years for blood. The overall multi-tissue accuracy rate based on bootstrap analysis was 7.8 years (95% Confidence Interval 6–9.7 years). The identified multi-tissue age prediction model has the potential to assist forensic investigations without the requirement to identify the sample body fluid type.

Keywords: DNA methylation, Epigenetics, HM450, Microarray, Aging, CpG sites.

1 Introduction

As age estimation provides intelligence in a forensic investigation, a considerable number of studies have identified age related (AR) CpG markers in various tissues and have developed age prediction assays applicable to body fluids of forensic interest. However, using the methylation level of one set of tissue-specific AR markers to predict the chronological age of a sample donor from other tissues can result in poor prediction accuracy [1]. Therefore, identifying a universal set of AR CpG markers to predict an individual's chronological age will lower the estimation error as well as bypass the necessity to first identify the tissue type, a step that often exposes the valuable DNA evidence to chemical destruction.

In this study, we aimed to identify a set of universal AR CpG biomarkers that are common between body fluids frequently found at crime scenes (blood, saliva, semen, menstrual blood, and vaginal secretions). Methylation data for these tissues will be used to develop a multi-tissue age prediction model with no more than 10 CpG markers. We have restricted the number of CpG markers to 10 or less to facilitate the design of PCR based DNA methylation assay that can be implemented in any forensic laboratory.

2 Materials and methods

2.1 Data

The training data were downloaded from the NCBI database (accession number GSE59509) [2]. The dataset consists of 41 samples, profiled on the Illumina Infinium HM450 platform. Donors ranged from 20 to 59 years old, with samples coming from five different body fluid types, namely blood, saliva, semen, menstrual blood, and vaginal secretions. The DNA methylation levels (beta values) at 450,000 CpG sites were normalized using beta mixture quantile dilation (BMIQ), and probes containing a known SNP marker or hybridising non-specifically to the human genome were removed from the dataset [3].

2.2 Statistical analysis

Elastic-net regression was performed using the *glmnet* package in R. The alpha parameter of glmnet was chosen to be 0.5. The number of features used by the model can be controlled by the shrinkage parameter lambda. This was adjusted to build models using up to 10 CpG markers with cross-validation.

2.3 Validating the prediction models

The testing data consists of 24 samples (three menstrual blood, three vaginal secretions, eight buccal swabs, and ten blood samples). Samples were collected from three different datasets downloaded from the NCBI repository. The DNA methylation profiles in the validation data were

assessed and normalized using BMIQ, and were then compiled into one data set along with their chronological ages ranged from 19 to 61 years old.

3 Results and discussion

3.1 Multi-tissue age prediction model

Elastic net regression was used to derive predictive models, whilst selecting features by using the shrinkage parameter. Fifty four AR CpG sites were selected across the tissues from a starting set of 310,014. Among these markers, a model with ten sites was selected based on its performance on the training data set. Maintaining a minimum number of CpG markers in the predictive model makes it possible to develop a multiplex PCR test that is compatible with typical forensic science laboratory equipment, rapid to analyse and has the ability to work on small/degraded DNA samples, which are critical features of any forensic test. Interestingly, only two of these ten CpG sites have been previously identified in the literature as being associated with age. The prediction accuracy of the model was calculated as the mean absolute deviation (MAD) between predicted and chronological age, based on the training data set, and was equal to 3.8 years (**Fig. 1A**).

3.2 Model validation

Based on the validation dataset, the MAD value for the menstrual blood and vaginal fluid is 6.9 years, for buccal swabs is 5.6 years, and 7.8 years for blood samples. The overall multi-tissue accuracy rate based on bootstrap analysis is 7.8 years with 95% confidence intervals of 6 to 9.7 years. The Pearson's correlation coefficient (r) between the actual and predicted age is 0.73 ($p<0.05$) (**Fig. 1B**). Finally, it is important to note that, as with all models of this type, the predictive model constructed here is only applicable to samples taken from the same tissue types, and within the age range of those samples used to train the model.

4 Conclusion

The purpose of this work was to use published datasets to screen the epigenome, to identify a small sub-set of CpG markers for age estimation across forensically relevant tissues which could potentially be incorporated into a multiplex PCR assay. This study has identified a multi-tissue predictive model based on ten CpG markers selected by elastic net regression. This model displayed a high predictive capability on the multi-tissue samples from the test data set, with a prediction accuracy of 7.8 years. This suggests that the selected epigenetic markers could be used for age estimation using blood, semen, saliva, menstrual blood, and vaginal secretions as a biological substrate. In future investigations, this model may be further improved with the inclusion of a larger number of samples with a broader age spectrum.

Conflict of interest statement

Daniel Halligan is an employee of Fios Genomics Ltd, a contract research organisation that provide bioinformatics services, however this work was undertaken voluntarily and without payment.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Hussain Alsaleh was supported by doctoral funding awarded by the Ministry of Interior of Kuwait.

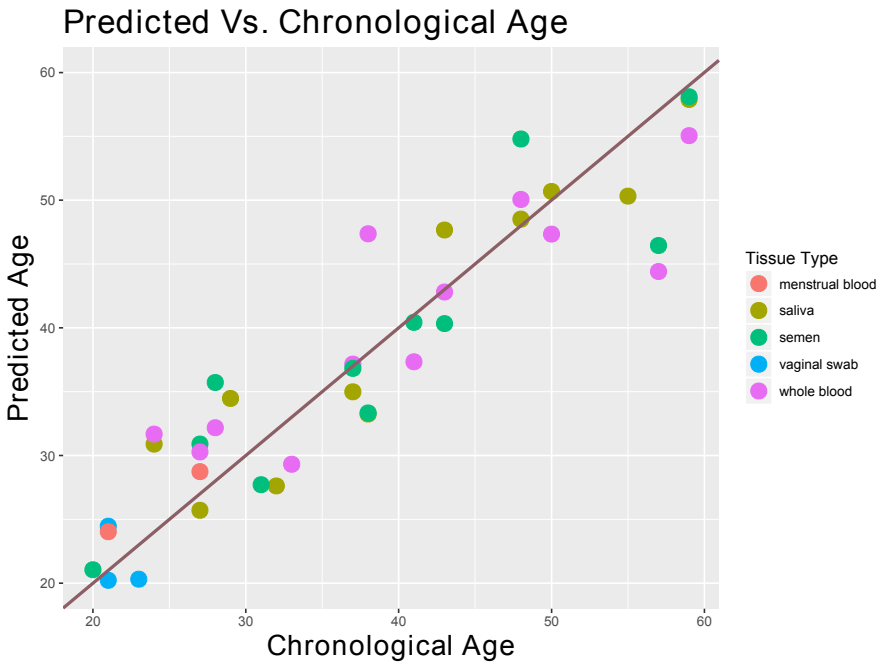
Reference:

[1] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sánchez, J.S. Sinsheimer, S. Horvath, et al., Epigenetic Predictor of Age, Plos One. 6 (2011) e14821.

[2] J.H. An, S.-E. Jung, Y.N. Oh, E.Y. Lee, A. Choi, W.I. Yang, et al., Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers, Forensic Science International: Genetics. 17 (2015) 17–24.

[3] E.M. Price, A.M. Cotton, L.L. Lam, P. Farré, E. Emberly, C.J. Brown, et al., Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array, Epigenetics & Chromatin 2013 6:1. 6 (2013) 1.

A



B

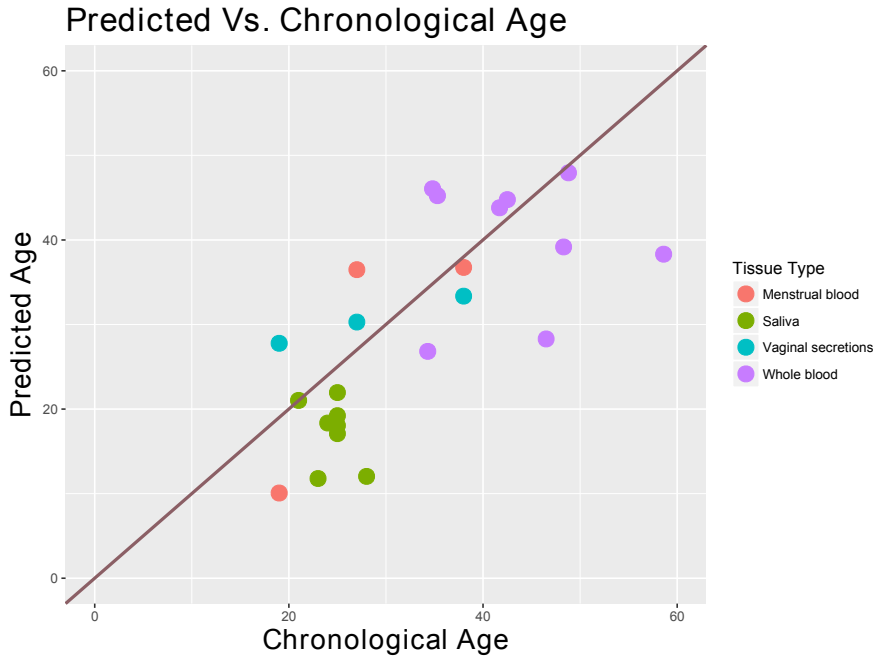


Fig. 1. Age prediction performance on the training (A) and validation dataset (B), based on the ten CpG markers elastic net model.