INFERENCE OF AGE PREDICTION ON TWIN SAMPLES BASED ON DNA METHYLATION USING AGENA BIOSCIENCE (SEQUENOM) EPITYPER

A. Freire-Aradas[1*], C. Phillips[1], A. Mosquera-Miguel[1], L. Girón-Santamaría[1], A. Gómez-Tato[2], J. Álvarez-Dios[2], M. Casares de Cal[2], E. Pośpiech[3], W. Branicki[4], A. Parys-Proszek[5], Á. Carracedo[1], M.V. Lareu[1]

INTRODUCTION:

Individual age estimation has the potential to provide key information that could enhance and extend DNA intelligence tools. Following predictive tests for externally visible characteristics developed in recent years, prediction of age could contribute to the following forensic analysis: i. to guide police investigations in the absence of eyewitness testimony or a national DNA database entry; ii. to help in the analysis of unidentified human remains; iii. to contribute to individual age estimation in legal disputes; and iv. to improve age-associated phenotype prediction (such as hair colour and early onset of male pattern baldness).

DNA methylation at CpG positions has emerged as the most promising DNA tests to ascertain the individual age of the donor of a biological contact trace. Using the technology Agena Bioscience EpiTYPER, we have previously developed and validated an age prediction model based on blood samples from 725 European individuals at a total of seven high age-correlated DNA methylation markers using a multivariate quantile regression analysis [1]. Here we report the predictive performance of the developed model using a testing set of 52 monozygotic twin pairs assessed in the open-access *Snipper* forensic classification website.

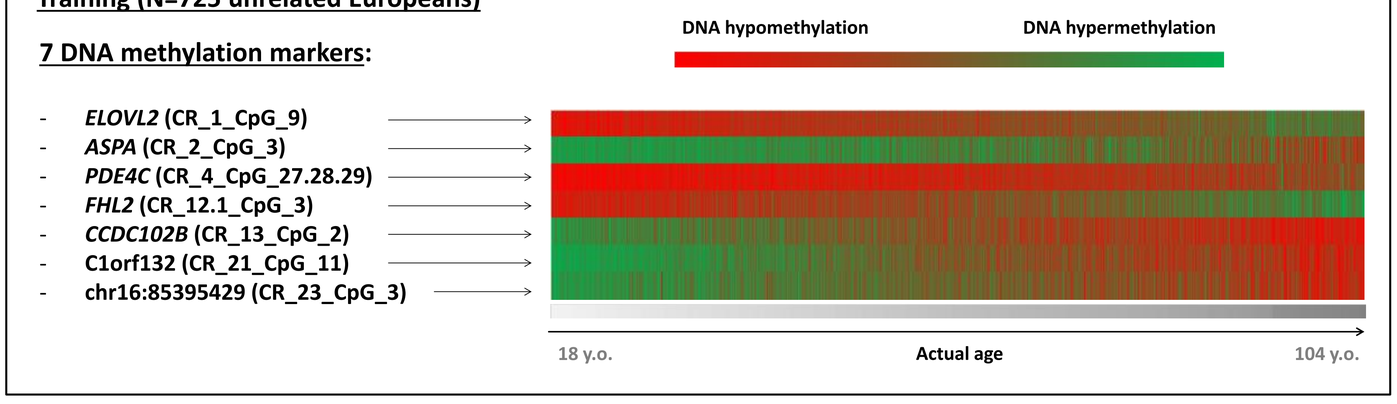


Fig. 1: Raster plot of the seven CpG sites from the previously developed age prediction model. A total of 725 European individuals are represented from left to right ordered by chronological age (18 to 104 years old) . Levels of DNA hypo- and hypermethylation are depicted in red and red green schemes.

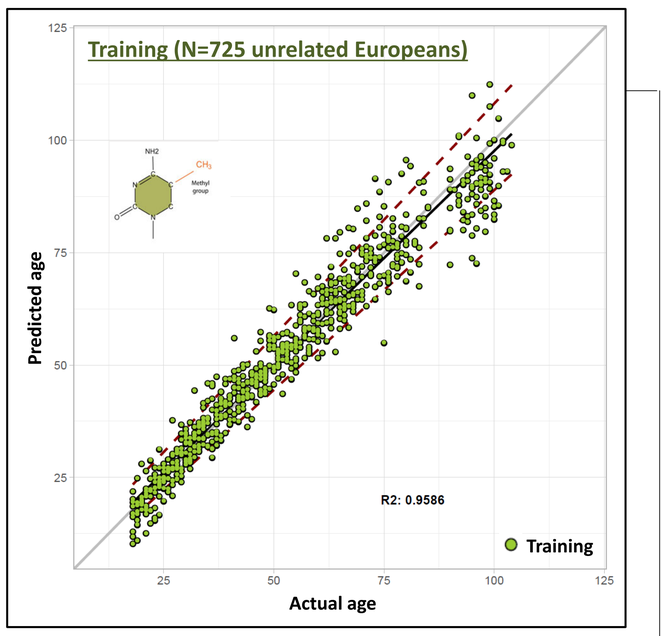


Fig. 2: Predicted versus actual age for the previously developed age prediction model using a training set of 725 European individuals from 18 to 104 years old (green data points) based on multivariate quantile regression analysis. The black diagonal line represents the 0.5 quantile regression line between predicted and actual ages and the discontinuous (dark red) lines, the corresponding 0.1 and 0.9 quantile regression lines (prediction intervals). The grey line is the diagonal line representing perfect correlation.

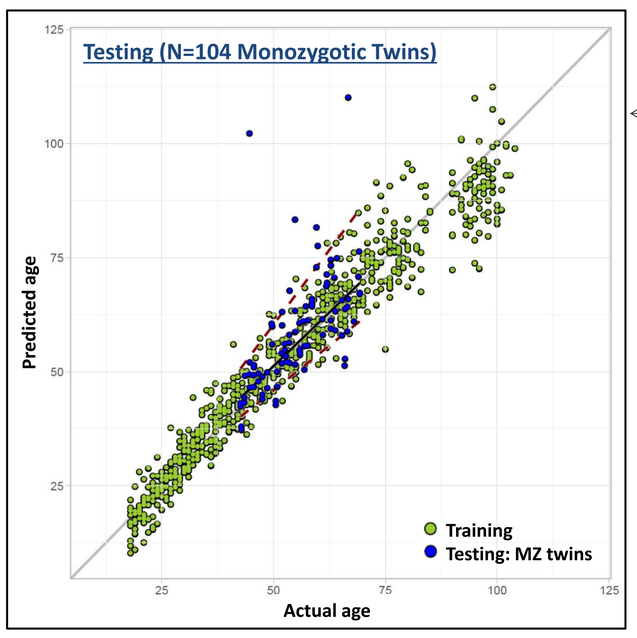


Fig. 3: Predicted versus actual age for the 104 monozygotic twins ranging from 42 to 69 years old (blue data points) using the above age prediction model. The black diagonal line represents the 0.5 quantile regression line between predicted and actual ages and the discontinuous (dark red) lines, the corresponding 0.1 and 0.9 quantile regression lines for the testing twins. The grey line is the diagonal line representing perfect correlation.

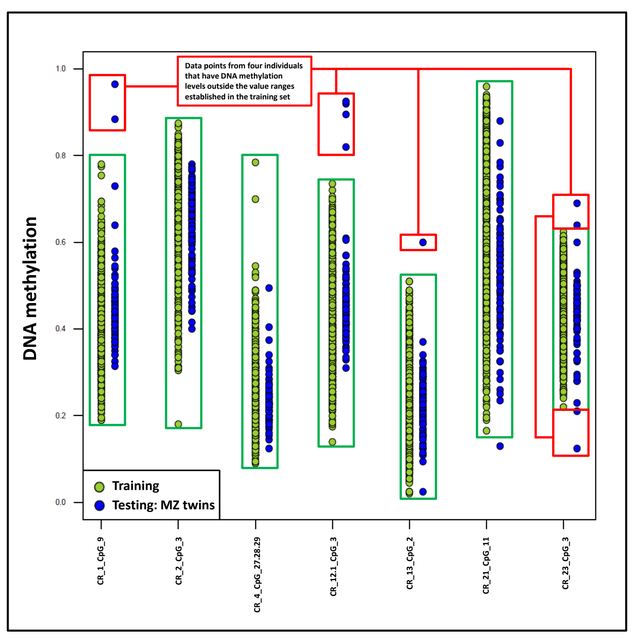
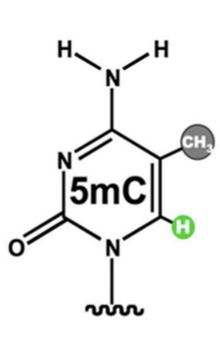


Fig. 4: Plot depicting the amplitude of DNA methylation values in training set samples (green points) and the monozygotic twin test set (blue points) in the seven CpGs used in the prediction model. Red squares highlight data points from four individuals amongst the 52 twin pairs analyzed that have DNA methylation levels outside the value ranges established in the training set.



MATERIAL & METHODS:

Peripheral blood samples from a total of 52 female monozygotic twin pairs (N=104) ranging from 42 to 69 years old were obtained from the Murcia Twin Registry, University of Murcia, Spain. Samples were bisulfite converted and DNA methylation was detected using Agena Bioscience (formerly Sequenom) following manufacture´s guidelines [2]. Target CpG sites (DNA methylation markers) were: *ELOVL2* (CR_1_CpG_9), *ASPA* (CR_2_CpG_3), *PDE4C* (CR_4_CpG_27.28.29), *FHL2* (CR_12.1_CpG_3), *CCDC102B* (CR_13_CpG_2), C1orf132 (CR_21_CpG_11) and chr16:85395429 (CR_23_CpG_3) (**Fig. 1**). Age predictions were performed using ***Snipper*** Forensic Classifier [3] based on a previously developed training set of 725 European individuals ranging from 18 to 104 years old (**Fig. 2**).

RESULTS – Part I – Multivariate quantile regression:

A group of 104 female monozygotic twin samples (52 pairs, 42-69 years old) was used to test the age prediction model based on a multivariate quantile regression analysis described in [1]. No results were obtained from twelve samples due to missing data, with the current prediction model only accepting complete methylation data. The other 46 pairs produced 83.70% correct predictions (actual age inside the prediction intervals) with a median absolute prediction error of ± 4.23 **(Fig. 3).** The higher error found in this sample group was mainly explained by a small number of outlier individuals with predicted ages highly divergent from their actual age, comprising four donors with more than 20 years between predicted and actual age. Methylation signals for these individuals were further assessed and it was observed that in some DNA methylation markers, the corresponding values lay outside the methylation range established in the training set **(Fig. 4).** For the small proportion of individuals with methylation values outside those established for the prediction model, the online age predictor has been constructed to display a warning in such situations and halt calculations, to avoid false predictions. Twin concordance was analyzed and gave a median difference of ± 3.41 years estimated age within pairs.

Fig. 5 illustrates the analysis of example epigrams and mass spectra obtained from EpiTYPER and subsequent statistical output from ***Snipper*** in a 43-year-old twin pair (twin donor A and B). Epigrams correspond to the amplicon *ELOVL2* (CR_1; 362 bp) covering a total of thirteen CpG sites (coloured circles on the horizontal line), including the most age-correlated CpG of CR_1_CpG_9. Mass spectra for CR_1_CpG_9 are outlined and either unmethylated (red) or methylated (blue) signals are highlighted for both donors and indicate similar DNA methylation ratios. Subsequent analysis using *Snipper* of methylation levels in the seven markers gives the predicted age and intervals. The twin donor A is predicted to be 43.12 years old with prediction intervals of 37.07-49.44 years. The twin donor B is predicted to be 42.37-year-old with prediction intervals of 34.55-48.12 years.

Fig. 6 displays the amplitude of the prediction errors for the twin samples. From each twin pair, donors A and B are depicted in orange and blue respectively. Both patterns of over- and underestimation of the actual age can be observed.

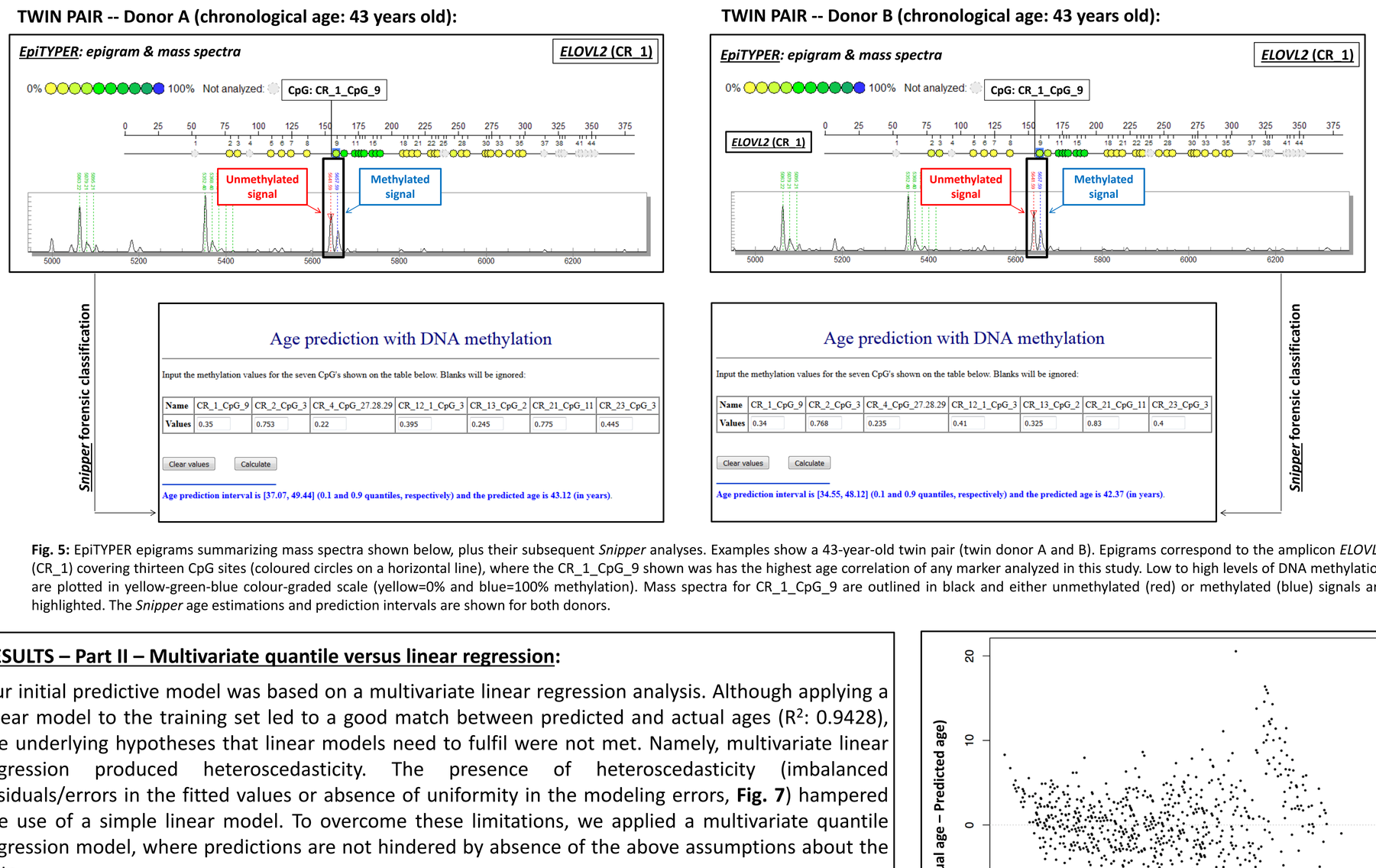


Fig. 5: EpiTYPER epigrams summarizing mass spectra shown below, plus their subsequent *Snipper* analyses. Examples show a 43-year-old twin pair (twin donor A and B). Epigrams correspond to the amplicon *ELOVL2* (CR_1) covering thirteen CpG sites (coloured circles on a horizontal line), where the CR_1_CpG_9 shown was has the highest age correlation of any marker analyzed in this study. Low to high levels of DNA methylation are plotted in yellow-green-blue colour-graded scale (yellow=0% and blue=100% methylation). Mass spectra for CR_1_CpG_9 are outlined in black and either unmethylated (red) or methylated (blue) signals are highlighted. The *Snipper* age estimations and prediction intervals are shown for both donors.

RESULTS – Part II – Multivariate quantile versus linear regression:

Our initial predictive model was based on a multivariate linear regression analysis. Although applying a linear model to the training set led to a good match between predicted and actual ages (R^2: 0.9428), the underlying hypotheses that linear models need to fulfil were not met. Namely, multivariate linear regression produced heteroscedasticity. The presence of heteroscedasticity (imbalanced residuals/errors in the fitted values or absence of uniformity in the modeling errors, **Fig. 7**) hampered the use of a simple linear model. To overcome these limitations, we applied a multivariate quantile regression model, where predictions are not hindered by absence of the above assumptions about the data.

Additionally, quantile regression establishes age-specific prediction intervals each time new data contribute to the model, with the condition that the new methylation signals are in the range of those of the training set. If methylation values are out of this range, as occurred in four of our analyzed twin samples, extrapolation of the data is not appropriate for quantile or linear regression and false predictions can occur. To highlight this limitation, the online age prediction tool will show warnings accordingly and no age prediction interval will be given in such cases.

Fig. 7: Presence of heteroscedasticity plotted as the residuals (actual age – predicted age) versus the predicted age for the 725 European samples. (training set)

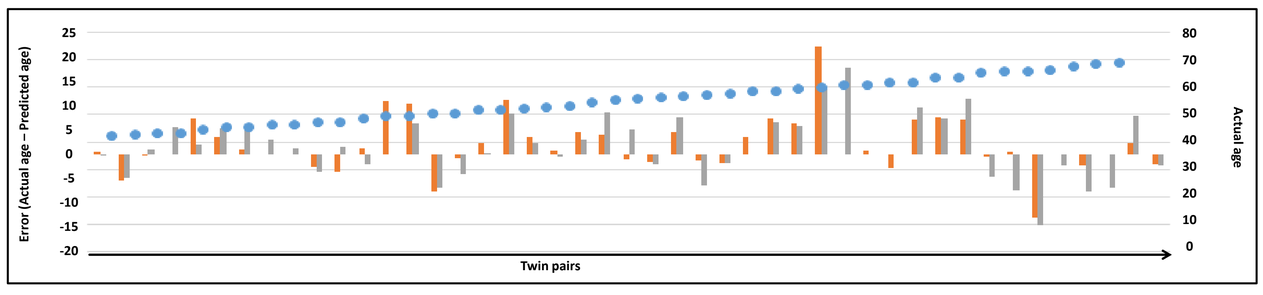


Fig. 6: Plot depicting the amplitude of the prediction errors for the twin samples. From each twin pair (x-axis), donors A and B are depicted in orange and blue respectively. Left y-axis represents the error in prediction (actual age – predicted age) and the right y-axis represents the actual age of the twin samples (blue points).

REFERENCES:

[1] A. Freire-Aradas et al, Development of methylation marker sets for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system, Manuscript submitted.
[2] M. Ehrich et al, Introduction to EpiTYPER for quantitative DNA methylation analysis using the MassARRAY® system, Sequenom® Application Note (2006).
[3] http://mathgene.usc.es/cgi-bin/snps/processmethylation.cgi

Filiations:

[1] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain.
** contact email: ana.freire@usc.es*
[2] Faculty of Mathematics, University of Santiago de Compostela, Spain.
[3] Institute of Zoology, Faculty of Biology and Earth Sciences, Jagiellonian University, Kraków, Poland.
[4] Faculty of Biochemistry, Biophysics and Biotechnology of the Jagiellonian University, Krakow, Poland
[5] Section of Forensic Genetics, Institute of Forensic Research, Kraków, Poland.

Acknowledgements:

AFA was supported by post-doctorate funding awarded by the Xunta de Galicia, Spain (as part of the Plan Galego de Investigación, Innovación e Crecemento 2011–2015, Axudas de apoio á etapa de formación postdoutoral, Plan I2C). MVL was supported by the Ministry of Economy and Competitiveness, Spain (BIO2013-42188-R). AM received financial support from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 285487 (EUROFORGEN-NoE). The National DNA Bank Carlos III is supported by ISCIII, Ministry of Science and Innovation, Spain (PT13/0001/0037, PT13/0010/0067): The Murcia Twin Registry is supported by the Seneca Foundation, Regional Agency for Science and Technology, Murcia, Spain (15302/PHCS/10) and Ministry of Science and Innovation, Spain (PSI11560-2009). The genotyping service generating methylation analysis data for this study was carried out at CEGEN-PRB2-ISCIII and is supported by grant PT13/0001, ISCIII-SGEFI / FEDER. .