



UNIVERSIDADE D
COIMBRA

Maria Helena Correia Dias

**DNA METHYLATION AS AN AGE PREDICTOR
IN LIVING AND DECEASED INDIVIDUALS**

Tese no âmbito do Doutoramento em Antropologia, Antropologia Forense, orientada pelos Professores Doutores Licínio Manuel Mendes Manco, Eugénia Maria Guedes Pinto Antunes da Cunha, Francisco Manuel Andrade Corte Real Gonçalves e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Abril de 2021

PhD in Anthropology, Forensic Anthropology

Faculty of Sciences and Technology

University of Coimbra

Title:

DNA methylation as an age predictor in living and deceased individuals

(Metilação de DNA como um preditor de idade em indivíduos vivos e mortos)

Author:

Maria Helena Correia Dias

Supervisors:

Licínio Manuel Mendes Manco^a,

Eugénia Maria Guedes Pinto Antunes da Cunha^{b,c},

Francisco Manuel Andrade Corte Real Gonçalves^{c,d}

Affiliations:

^a University of Coimbra, Research Centre for Anthropology and Health, Department of Life Sciences, Portugal

^b University of Coimbra, Centre for Functional Ecology, Department of Life Sciences, Portugal

^c National Institute of Legal Medicine and Forensic Sciences, Portugal

^d Faculty of Medicine of University of Coimbra, Portugal

Date: April 2021

Funding: HCD has a PhD grant from FCT (SFRH/BD/117022/2016).



REPÚBLICA PORTUGUESA **CENTRO** 2020



UNIÃO EUROPEIA

Fundo Social Europeu

Acknowledgments/Agradecimentos

O meu trabalho de doutoramento teve o apoio e colaboração de várias entidades e pessoas pelo que não poderia estar mais grata por todo o auxílio e disponibilidade prestada durante a realização do mesmo.

Ao meu orientador, **Doutor Licínio Manco**, por ter aceitado este desafio comigo, por toda a paciência, compreensão e por toda a ajuda diária e constante. Por todas as trocas de ideias, por toda a transmissão de conhecimentos científicos e toda a orientação de qualidade que permitiu o meu crescimento académica e pessoalmente. O meu eterno agradecimento por toda a amizade e confiança, por estar sempre presente quando eu precisava, e por ter sempre tentado perceber os meus pontos de vista.

À minha coorientadora, **Doutora Eugénia Cunha**, por todas as palavras de carinho e compreensão. Por ser além de Professora e guia em termos académicos, uma amiga presente e confidente. A sua notória orientação científica e a sua clareza na transmissão de conhecimentos contribuíram para o sucesso do meu trabalho. O meu agradecimento do fundo do coração por me ter aceitado como sua orientanda, por tudo o que fez por mim, por estar sempre lá, e por me ter ajudado a crescer pessoal e profissionalmente.

Ao meu coorientador, **Doutor Francisco Corte Real**, por ter aceitado a orientação do meu trabalho de doutoramento e por todo o apoio e disponibilidade durante a realização do mesmo. O meu eterno agradecimento por toda a ajuda na resolução de

Agradecimentos

problemas burocráticos de maior dimensão e, essencialmente, obrigada por toda a confiança depositada, todo o incentivo e força quando um resultado não era o esperado, ou um artigo não era aceite.

Às minhas instituições de acolhimento **CIAS (LGH)**, **INMLCF** e **CEF (LAF)**, por terem reunido todas as condições necessárias para a realização deste trabalho de doutoramento e especialmente à **Professora Doutora Cristina Padez** e ao **CIAS** por todo o apoio prestado ao trabalho laboratorial.

À minha entidade financiadora **Fundação para a Ciência e Tecnologia (FCT)**, por ter permitido a realização deste trabalho de doutoramento através do contrato de concessão de bolsa. Sem dúvida, a realização deste trabalho e a obtenção deste grau académico não teria sido possível sem o apoio da FCT.

À **Dra. Cristina Cordeiro** pela ajuda incansável, pelo carinho e por todas as chamadas em que apenas mais uma amostra havia, mas era o suficiente, o necessário, para tornar mais rápido o que era, por si só, bastante demorado. Por toda a amizade e força que me deu ao longo do meu trajecto, agradeço-lhe do fundo do coração.

Aos **técnicos de autópsias do INMLCF das Delegações do Centro e Sul** pela disponibilidade na recolha das amostras, com a supervisão da **Dra. Cristina Cordeiro** e do **Dr. Carlos dos Santos**.

À **Professora Doutora Teresa Ferreira**, por toda a amizade, disponibilidade, ajuda, boa disposição e por todo o carinho durante o meu trabalho de doutoramento.

Agradecimentos

A todos os **Doutores do CEF, do CIAS e do Departamento de Ciências da Vida** por toda a ajuda ao longo de todos os anos da minha formação e por toda a disponibilidade sempre que precisava.

À **Dra. Lisa Sampaio** pela simpatia, alegria e amizade. Pela sua preocupação constante e ajuda no desempenho do meu trabalho, por todos os esforços que fez por mim e por todas as boas notícias que me deu.

Ao **Dr. Armando Serra**, a quem devo tudo o que sei relativamente ao processamento de restos humanos, por ter sido atencioso, compreensivo e sempre preocupado ao longo do meu trabalho.

A todos os **técnicos do Serviço de Génética Forense da Delegação do Centro do INMLCF** por terem conciliado comigo os seus horários, a fim de conseguirmos rentabilizar as horas em que poderia trabalhar no serviço.

À **Professora Doutora Helena Teixeira** pela ajuda nas questões burocráticas necessárias para agilizar os pedidos feitos ao INMLCF, pelo carinho, amizade, e disponibilidade ao longo destes anos.

Ao **Professor Doutor Duarte Nuno Vieira** por todo o apoio, amizade e disponibilidade ao longo do meu trajecto académico. Também lhe deixo o meu eterno agradecimento por todo o apoio prestado com o acesso aos Corpos Doados à Ciência.

Agradecimentos

À **Professora Doutora Dulce Madeira** e ao **Dr. Pedro Pereira** por todo o carinho, disponibilidade e preocupação constante comigo e com o meu trabalho e a **todos os técnicos do Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto** que ajudaram na recolha das amostras provenientes dos Corpos Doados à Ciência. Sem eles, parte desta investigação não teria sido possível.

Aos meus **pais** por todo o AMOR, carinho, paciência, compreensão de ausências e, principalmente, por estarem sempre a meu lado, desde os primórdios da minha existência até esta fase final do meu percurso académico. Apesar da distância, sempre estiveram mais presentes do que os presentes diários. Em especial “à **minha mais que tudo**”, a **minha Mãe**, um OBRIGADA por me teres acolhido nos teus braços desde o dia em que nasci, por me teres guiado para Coimbra, tua cidade e dos Doutores, por me teres feito amar esta cidade como tu a amavas e como Minha segunda Cidade. Por seres o meu elo protector, a minha paixão, a minha Santa que me guia e me ajuda a voltar para o porto seguro depois da tempestade, seja ela qual for. Por ti MÃE vim, por ti fiquei, e por ti sei que sempre irei mais longe. A ti dedico toda a essência deste trabalho.

Ao meu **Pakiko** por todos os sorrisos e abraços, por toda a meiguice de criança com a inocência que lhe é inerente. Por todos os “tu consegues tudo Titittita”. À minha **Miana** que me enche o coração com o seu sorriso meigo e amoroso e traz sempre muita alegria à Titi. Obrigada por serem um pilar gigante na minha vida.

À minha **irmã** por eu ser a sua eterna “Poppy”, e ao meu **irmão** por ser o meu segundo pai e por estar sempre presente e preocupado com a sua “Piquinina”, apesar da distância.

Agradecimentos

Às **minhas meninas de sempre e aos meus verdadeiros amigos de Coimbra** que me acompanharam durante toda a minha evolução pessoal e académica, um obrigada sincero por estarem sempre lá para tudo. Em especial, à minha **Catarina**, amizade inesperada mas verdadeira, por ter sido um apoio pessoal e profissional em todos os momentos. Obrigada por ter em ti uma das melhores pessoas de Coimbra.

Aos **meus amigos do “Lab”**, por toda a amizade e companherismo, por todas as ajudas estatísticas, por todas as trocas de ideias e pela boa disposição e carinho.

À **Senhora Vitória**, companhia de horas boas e más, por ser uma amizade sem idade e sem limites.

Ao meu **AMOR da Vida toda**, que entrou sem querer e decidiu ficar, que me aturou e me acarinhou. Que fazia com que os problemas se tornassem mais fáceis nos dias maus. Ao meu amor, por Coimbra te ter trazido, mas por Coimbra não te ter levado de mim. Obrigada por valeres a pena.

Aos **meus avós**, força omnipresente e constante, por terem acreditado em mim quando ainda nem sabia o que queria. Um obrigada especial à minha **Voinha**, que apesar de já não estar entre nós, nos dias maus era a ti que eu me agarrava para ter forças. Sei que estejas onde estiveres, estarás muito orgulhosa de mim.

Por fim, **a Ti**, peço desculpa por me ter perdido em tempos, mas a vida tem razões que a própria razão desconhece. Fizeste com que me encontrasse de novo e ensinaste-me que a Vida ainda valia a pena. Muito havia para fazer e tinha uma segunda oportunidade...

Agradecimentos

Agradeço com toda a minha essência que tenhas “aberto o teu manto” para eu não cair e me tenhas feito voltar pra Ti!

Index Contents

Abstract	I
Resumo	V
List of Abbreviations	IX
List of Tables	XIII
List of Figures	XXI

Chapter 1. General introduction

1

1. Forensic anthropology and forensic genetics	3
2. Age estimation in forensic contexts	5
2.1. Anthropological age estimation methods	8
2.2. Molecular and biochemical age estimation methods	12
3. Epigenetic age estimation	16
3.1. Evolution of epigenetics	16
3.2. DNA methylation (DNAm)	17
3.2.1. DNAm changes with age	19
3.2.2. Approaches used for the development of APMs based on DNAm	22
3.2.3. A brief review on DNAm age prediction models (APMs)	25
4. Challenges for application of DNAm age estimation in forensic contexts	44
4.1. Differences between predicted and chronological ages with aging	45
4.2. Specificity of each age-correlated marker to diseases	45
4.3. Tissue specificity	46
4.4. Specificity of population groups	47
4.5. Sex differences	48

Chapter 2. Objectives

49

Chapter 3. Sample and design research	55
1. Types of samples, collection, handling and storage	57
1.1. Blood samples	57
1.2. Bone samples	59
1.3. Tooth samples	61
1.4. Buccal swabs	61
2. Ethical or institutional guidelines	62
3. Laboratory methods	63
3.1. DNA extraction	63
3.1.1. Blood samples	63
3.1.2. Bone and tooth samples	63
3.1.3. Buccal swabs	66
3.2. DNA quantification	67
3.2.1. Blood samples and buccal swabs	67
3.2.2. Bone and tooth samples	67
3.3. Bisulfite conversion	69
3.3.1. Efficiency of bisulfite conversion	71
3.4. Polymerase Chain Reaction (PCR)	72
3.5. Sanger sequencing	73
3.5.1. Methylation quantification of CpGs after PCR direct DNA Sanger sequencing	75
3.5.2. Reproducibility of direct bisulfite sequencing	79
3.5.3. DNAm standards	79
3.6. Multiplex-PCR SNaPshot assay	79
3.6.1. Methylation quantification of CpGs after multiplex SNaPshot assay	86
3.7. Statistical analyses	87

Chapter 4. Results and discussion	91
A. DNA methylation age estimation in blood samples	93
1. Introduction	95
2. Materials and Methods	98
2.1. Sample collection	98
2.2. DNA extraction, quantification and bisulfite conversion	98
2.3. Polymerase chain reaction (PCR) and Sanger sequencing	99
2.4. SNaPshot assay	99
2.5. DNAm quantification	100
2.6. Statistical analyses	100
3. Results	102
3.1. Age estimation in blood samples from living individuals	103
3.1.1. DNAm data obtained in blood samples from living individuals using bisulfite Sanger sequencing	103
3.1.2. DNAm data obtained in blood samples from living individuals using SNaPshot methodology	112
3.1.3. Comparison between methodologies	119
3.2. Age estimation in blood samples from deceased individuals	119
3.2.1. DNAm data obtained in blood samples from deceased individuals using bisulfite Sanger sequencing	120
3.2.2. DNAm data obtained in blood samples from deceased individuals using SNaPshot methodology	127
3.2.3. Comparison between methodologies	132
3.3. Age estimation in blood samples from living and deceased individuals	133
3.3.1. DNAm data obtained in blood samples from living and deceased individuals using bisulfite Sanger sequencing	134
3.3.2. DNAm data obtained in blood samples from living and deceased individuals using SNaPshot methodology	140

3.3.3. Comparison between methodologies -----	146
3.4. Applicability of the developed APMs for blood samples from living and deceased individuals -----	147
3.4.1. Applying the developed APMs in blood samples from living individuals to deceased individuals -----	147
3.4.2. Applying the developed APMs in blood samples from deceased individuals to living individuals -----	148
4. Discussion -----	152
B. DNA methylation age estimation in tooth samples -----	167
1. Introduction -----	169
2. Materials and Methods -----	170
2.1. Sample collection -----	170
2.2. DNA extraction, quantification and bisulfite conversion -----	170
2.3. Polymerase chain reaction (PCR) and Sanger sequencing -----	171
2.4. SNaPshot assay-----	171
2.5. DNAm quantification -----	171
2.6. Statistical analyses -----	172
3. Results -----	173
3.1. DNAm data obtained in teeth from living and deceased individuals using bisulfite Sanger sequencing -----	173
3.2. DNAm data obtained in teeth from living and deceased individuals using SNaPshot methodology-----	178
3.3. Comparison between methodologies -----	183
4. Discussion -----	184

C. DNA methylation age estimation in fresh bone samples--189

1. Introduction	191
2. Materials and Methods	193
2.1. Sample collection	193
2.2. DNA extraction, quantification and bisulfite conversion	193
2.3. Polymerase chain reaction (PCR) and Sanger sequencing	194
2.4. SNaPshot assay	194
2.5. DNAm quantification	194
2.6. Statistical analyses	195
3. Results	196
3.1. Age estimation in bones collected during autopsies	196
3.1.1. DNAm data obtained in bones from autopsies using bisulfite Sanger sequencing	196
3.1.2. DNAm data obtained in bones from autopsies using SNaPshot methodology	201
3.1.3. Comparison between methodologies	204
3.2. Age estimation in bones collected from BDS	205
3.2.1. DNAm data obtained in bones from BDS using bisulfite Sanger sequencing	205
3.2.2. DNAm data obtained in bones from BDS using SNaPshot methodology	210
3.2.3. Comparison between methodologies	212
3.2.4. Evaluation of the effect of PMI (<i>postmortem</i> interval) in BDS	212
4. Discussion	214

D. DNA methylation age estimation in buccal swabs -----	221
1. Introduction -----	223
2. Materials and Methods -----	225
2.1. Sample collection -----	225
2.2. DNA extraction, quantification and bisulfite conversion -----	226
2.3. Polymerase chain reaction (PCR) and Sanger sequencing -----	226
2.4. SNaPshot assay -----	226
2.5. DNAm quantification -----	227
2.6. Statistical analyses -----	227
3. Results -----	228
3.1. DNAm data obtained in buccal swabs using bisulfite Sanger sequencing -----	229
3.2. DNAm data obtained in buccal swabs using SNaPshot methodology -----	235
3.3. Comparison between methodologies -----	241
4. Discussion -----	241
E. DNA methylation age estimation through multi-tissues ----	247
1. Introduction -----	249
2. Materials and Methods -----	251
2.1. Sample collection -----	251
2.2. DNAm analyses -----	251
2.3. Statistical analyses -----	251
3. Results -----	253
3.1. Multi-tissue APMs using Sanger sequencing -----	253
3.2. Multi-tissue APMs using SNaPshot methodology -----	265
4. Discussion -----	271

F. Dry bone samples from <i>Coleção de Esqueletos Identificados do Século XXI</i> - CEI/XXI	277
1. Introduction	279
2. Materials and Methods	280
2.1. Sample collection	280
2.2. Bone processing	281
2.3. DNA extraction, quantification and bisulfite conversion	283
2.4. Polymerase chain reaction (PCR) and Sanger sequencing	283
3. Results	286
4. Discussion	289
Chapter 5. General discussion	293
Chapter 6. Conclusion and Future directions	315
Chapter 7. Bibliography	321
Annex I	349
Forms (in Portuguese)	
Annex II	357
Biological features of Bodies Donated to Science (BDS)	
Annex III	361
Ethical consent (in Portuguese)	

Annex IV -----	365
Protocols	
Annex V -----	375
Supplementary Material Tables	
Annex VI -----	405
Supplementary Material Figures	
Annex VII -----	471
Description of bone samples collected from CEI/XXI (in Portuguese)	
Annex VIII -----	503
Published papers	

Abstract

Age estimation is very useful in forensic sciences for identification purposes. In recent years, DNA methylation (DNAm) has arisen as a promising tool for age prediction. In consequence, age prediction models (APMs) have been developed using many age-correlated genes, different methodologies and several biological samples.

The main aim of this study was to develop APMs based on previously known age-associated genes, including *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C*, *C1orf132*, *KLF14* and *TRIM59*, using different tissues with forensic interest. In addition, we investigated putative differences in DNAm between blood from living and deceased individuals and according to increasing age, sex, ancestry and *postmortem* interval (PMI).

DNAm levels were evaluated in 267 biological samples (including blood, teeth, bones and buccal swabs) from living and deceased Portuguese individuals through bisulfite polymerase chain reaction (PCR) Sanger sequencing and multiplex SNaPshot methodologies. Additionally, nine skeletonized individuals were tested for DNAm analysis. Linear regression models were used to analyze relationships between DNAm and chronological age.

Promising results were obtained in our study. Using SNaPshot, the most accurate APM built in 56 blood samples from living individuals included three CpGs (in genes *ELOVL2*, *FHL2* and *C1orf132*), showing a Mean Absolute Deviation (MAD) between predicted and chronological ages of 4.25 years; in 59 deceased individuals the most accurate APM was built with four CpGs (in genes *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*) with a MAD of 5.36 years. The Sanger sequencing allowed to build accurate APMs in 53 blood samples from living individuals (MAD = 5.35 years), with the best

CpGs at *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, and in 47 blood samples from deceased individuals (MAD = 6.08 years), with the best CpGs at *ELOVL2*, *FHL2*, *C1orf132*, *EDARADD* and *PDE4C*. In the combined set of blood samples from living and deceased individuals, the best APMs were built with CpGs at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* (MAD = 4.97 years) using SNaPshot, and with CpGs at genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* (MAD = 6.21 years) through Sanger sequencing.

Moreover, our study highlighted the most promising APM in 29 bones from autopsies, using Sanger sequencing, built with six CpGs at *ELOVL2*, *EDARADD* and *C1orf132*, and revealing a MAD of 2.56 years. Through SNaPshot, the most accurate APM (MAD = 7.18 years) in 31 bones from autopsies was built with two CpGs in genes *FHL2* and *KLF14*.

Potential useful APMs were developed in the other tissues, including: in 21 bones from Bodies Donated to Science with two CpGs in genes *FHL2* and *PDE4C* (MAD = 4.67 years) using Sanger sequencing; in buccal swabs with *TRIM59* CpG using SNaPshot in 35 samples (MAD = 6.73 years) and with two CpGs at *ELOVL2* using Sanger sequencing in 23 samples (MAD = 8.32 years); in teeth with two CpGs at *ELOVL2* and *KLF14* (MAD = 7.07 years) in 23 samples using SNaPshot.

Additionally, we developed several multi-tissue APMs, obtaining the best accuracy (MAD = 6.49 years) in a group with bones from autopsies and blood and teeth from living and deceased individuals using SNaPshot (with CpGs at *ELOVL2*, *KLF14* and *C1orf132*). In this same group, Sanger sequencing allowed the development of an APM with nine CpGs at genes *ELOVL2*, *EDARADD*, *PDE4C* and *FHL2* and a similar value of MAD (6.42 years).

Ours results revealed that the model accuracy depends on the chronological age of individuals, in concordance with previous studies, and that there is no influence of sex

and PMI in DNAm levels. The possibility that age markers might be population specific and that *postmortem* changes can alter the methylation status among specific loci was also suggested by our data. DNAm changes were observed due to treatment of the body after death. In regards to bones from skeletonized individuals, the amount of extracted DNA was not sufficient for a subsequent successful DNAm age estimation.

Moreover, it has been observed that the APMs developed in blood from living individuals cannot be applied to deceased individuals with a similar MAD value. However, the APMs developed in blood from deceased individuals when applied to living individuals revealed a similar MAD. This can be a relevant issue in future application of the developed APMs in forensic contexts.

In conclusion, our study is a relevant contribution for DNAm age research, showing the usefulness of both bisulfite PCR sequencing and multiplex SNaPshot methodologies for forensic analysis in several tissue types. The herein developed APMs seem to be informative and could have potential applications in different forensic contexts. Despite of this, our study could be improved with the use of larger sample sets, with higher statistical power, and being more representative of age range and different tissue types.

Key words

Forensic Sciences, Age estimation, DNA methylation (DNAm), Age prediction models (APMs), Sanger sequencing, SNaPshot

Resumo

A estimativa da idade é muito útil em Ciências Forenses para fins de identificação. Nos últimos anos, a metilação de DNA (DNAm) surgiu como uma ferramenta promissora para predição da idade. Em consequência, vários modelos de predição de idade (APMs) têm sido desenvolvidos, recorrendo a genes correlacionados com a idade, diferentes metodologias e várias amostras biológicas.

O principal objetivo deste estudo foi desenvolver APMs usando genes previamente associados com a idade, incluindo *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C*, *C1orf132*, *KLF14* e *TRIM59*, e diferentes tecidos biológicos com interesse forense. Além disso, investigámos possíveis diferenças nos níveis de DNAm entre sangue de indivíduos vivos e mortos, e de acordo com as categorias etárias, sexo, ancestralidade e intervalo *postmortem* (PMI).

Os níveis de DNAm foram avaliados em 267 amostras biológicas (incluindo sangue, dentes, ossos e esfregaços bucais) de indivíduos Portugueses, vivos e mortos, através das metodologias de sequenciação de Sanger e SNaPshot, após conversão do DNA por bissulfito. Foram ainda testados 9 indivíduos esqueletizados para análise de DNAm. A associação entre os níveis de DNAm e a idade cronológica foi analisada com modelos de regressão linear.

Foram obtidos alguns resultados promissores, entre os quais se destacam os melhores APMs, obtidos por SNaPshot, em 56 amostras de sangue de indivíduos vivos e que incluiu 3 CpGs dos genes *ELOVL2*, *FHL2* e *C1orf132*, com uma média de desvio absoluto (MAD) entre as idades prevista e cronológica de 4,25 anos, e o melhor APM em 59 amostras de sangue de indivíduos mortos, construído com 4 CpGs dos genes *ELOVL2*,

FHL2, *C1orf132* e *TRIM59* e uma MAD de 5,36 anos. Por sua vez, por sequenciação, os melhores APMs foram obtidos em 53 amostras de sangue de indivíduos vivos com o melhor CpG de cada um dos genes *ELOVL2*, *FHL2*, *EDARADD* e *PDE4C* (MAD = 5,35 anos), e em 47 amostras de sangue de indivíduos mortos com o melhor CpG de cada um dos genes *ELOVL2*, *FHL2*, *C1orf132*, *EDARADD* e *PDE4C* (MAD = 6,08 anos). No conjunto de sangue de vivos e mortos, os melhores APMs foram obtidos com os CpGs dos genes *ELOVL2*, *FHL2*, *C1orf132* e *TRIM59* (MAD = 4,97 anos) por SNaPshot, e com 4 CpGs dos genes *ELOVL2*, *FHL2*, *EDARADD* e *PDE4C* (MAD = 6,21 anos) por sequenciação.

O nosso estudo destacou um dos APMs mais promissores em 29 ossos de autópsias usando sequenciação, construído com 6 CpGs dos genes *ELOVL2*, *EDARADD* e *C1orf132*, e uma MAD de 2,56 anos. Através de SNaPshot, o melhor APM foi construído em 31 ossos de autópsias com 2 CpGs dos genes *FHL2* e *KLF14* (MAD = 7,18 anos).

Outros APMs foram desenvolvidos nos restantes tecidos, incluindo: em 21 ossos dos Corpos Doados à Ciência, com 2 CpGs dos genes *FHL2* e *PDE4C* (MAD = 4,67 anos), através da sequenciação; em esfregaços bucais, com o CpG do gene *TRIM59* (MAD = 6,73 anos) em 53 amostras por SNaPshot, e com 2 CpGs do gene *ELOVL2* (MAD = 8,32 anos) em 23 amostras por sequenciação; em dentes, com 2 CpGs dos genes *ELOVL2* e *KLF14* (MAD = 7,07 anos) em 23 amostras, por SNaPshot.

Também desenvolvemos APMs em conjuntos de vários tecidos, tendo-se obtido a melhor precisão (MAD = 6,49 anos) para o APM construído no grupo com ossos de autópsias e sangue e dentes de indivíduos vivos e mortos, por SNaPshot (com 3 CpGs dos genes *ELOVL2*, *KLF14* e *C1orf132*). No mesmo grupo, por sequenciação, desenvolveu-

se um APM com 9 CpGs dos genes *ELOVL2*, *EDARADD*, *PDE4C* e *FHL2* e uma MAD de 6,42 anos.

Os nossos resultados revelaram que a precisão dos APMs depende da idade cronológica dos indivíduos, em concordância com estudos prévios, e que não há influência do sexo e do PMI nos níveis de DNAm. A possibilidade dos marcadores de idade serem específicos para as populações e de modificações *postmortem* poderem alterar o estado de metilação de certos genes foi também sugerida pelos nossos dados. Foram ainda observadas alterações nos níveis de DNAm devido ao tratamento do corpo após a morte. Em relação aos ossos de indivíduos esqueletizados, a quantidade de DNA extraído não foi suficiente para estimar a idade através de DNAm.

Foi ainda observado que os APMs desenvolvidos em sangue de vivos não podem ser aplicados a mortos com valores de MAD semelhantes. Mas, por sua vez, os APMs desenvolvidos em indivíduos mortos quando aplicados a vivos revelaram um valor de MAD similar. Isto pode ser importante na futura aplicação dos APMs em contextos forenses.

Em conclusão, o nosso estudo é uma contribuição relevante para a investigação da estimativa da idade através de DNAm, mostrando que as duas metodologias, sequenciação Sanger e SNaPshot, podem ser úteis para análises forenses em vários tecidos. Os APMs desenvolvidos parecem ser informativos e podem ter potencial aplicação em diferentes contextos forenses. Apesar disso, o nosso estudo poderia ser melhorado com o uso de um maior número de amostras, com maior poder estatístico, e maior representatividade quer de categorias etárias quer de diferentes tecidos.

Palavras-chave

Ciências Forenses, Estimativa da idade, Metilação de DNA (DNAm), Modelos de predição de idade (APMs), Sequenciação de Sanger, SNaPshot

List of Abbreviations

A, adenine

APM, Age Prediction Model

AGEs, advanced glycation end products

AAR, Aspartic acid Racemization

BDS, Bodies Donated to Science

CEI/XXI, *Coleção de Esqueletos Identificados do Século XXI*

CEF, *Centro de Ecologia Funcional*

Clorf132, chromosome 1 open reading frame 132; *MIR29B2CHG**

C, Cytosine

CIAS, *Centro de Investigação em Antropologia e Saúde*

CGIs, CpG islands

Chr, chromosome

CH₃, methyl group

5C, fifth carbon position

C^m or 5mC, methylated cytosine or 5-methylcytosine

CpG, cytosine-phosphate-guanine

DNAm, DNA methylation

DNMT, DNA methyltransferases

Exp_{CpG}, Expected CpG

EDARADD, EDAR Associated Death Domain*

ELOVL2, ELOVL Fatty Acid Elongase 2*

FA, Forensic anthropology

FG, Forensic Genetics

FHL2, Four And A Half LIM Domains 2*

FASE, Forensic Anthropology Society of Europe

Page | X

G, Guanine

GRCh38, Genome Reference Consortium Human Build 38

INMLCF, *Instituto Nacional de Medicina Legal e Ciências Forenses*

Interpol, International Criminal Police Organization

ISC/XXI, The 21st Century Identified Skeletal Collection

LGH, *Laboratório de Genética Humana*

LAF, *Laboratório de Antropologia Forense*

MAD, mean absolute deviation

MAE, mean absolute error

MPS, massively parallel sequencing

MS-HRM, methylation-sensitive high resolution melting

Min, minute

mtDNA, Mitochondrial DNA

N, number of samples

KLF14, Kruppel like factor 14*

Obs_{CpG}, Observed CpG

PCR, Polymerase Chain Reaction,

PMI, *Postmortem* interval

PDE4C, phosphodiesterase 4C, cAMP specific*

R, correlation coefficient beta, β ; Pearson correlation

r, Spearman correlation coefficient

RENDA, *Registo Nacional de Não Dadores*

RMSE, Root mean square error

SE, standard error

sjTREC, Signal joint T-cell receptor excision circles

STR, short tandem repeats

Sec, seconds

SGB, *Serviço de Genética e Biologia Forenses*

SBE, Single-base extension

T, Thymine

TRIM59, tripartite motif containing 59*

U, Uracil

vs., versus

*HGNC- *Hugo Gene Nomenclature Committee*

List of Tables

Chapter 1. General introduction

Table 1.1: Molecular and biochemical methods used in age predictions.

Table 1.2: Review of studies based on DNAm age estimation.

Chapter 2. Objectives

Chapter 3. Sample and design research

Table 3.1: Volume of each reagent per sample.

Table 3.2: PCR primers and sequence to analyse for PCR-sequencing.

Table 3.3: Multiplex PCR primers adapted from Jung *et al.* (2019). The concentrations of *ELOVL2* and *TRIM59* genes were modified in the multiplex amplification SNaPshot reaction, to obtain a clear and strong band from these genes.

Table 3.4: Sequencing primers used for the multiplex methylation SNaPshot assay, according to Jung *et al.* (2019).

Chapter 4. Results and discussion

A. DNA methylation age estimation in blood samples

Table 4.1: Comparison of two regression lines between males and females in blood samples from living individuals using data obtained from Sanger sequencing.

Table 4.2: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in blood samples from living individuals.

Table 4.3: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* in blood samples from living individuals.

Table 4.4: MAD between predicted and chronological ages stratified by age group in the training set of 53 blood samples from living individuals.

Table 4.5: MAD between predicted and chronological ages stratified by age group in the test set of 18 blood samples from living individuals.

Table 4.6: Comparison of two regression lines between males and females in blood samples from living individuals using data obtained from SNaPshot methodology.

Table 4.7: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using a SNaPshot assay in 59 blood samples from living individuals.

Table 4.8: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *FHL2* and *C1orf132*, selected by stepwise regression approach, in blood samples from living individuals.

Table 4.9: MAD between predicted and chronological ages stratified by age group in the training set of 56 blood samples from living individuals.

Table 4.10: Comparison of age-correlated values obtained in blood samples from living Portuguese individuals through Sanger and SNaPshot methodologies.

Table 4.11: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between DNAm levels obtained by bisulfite sequencing and chronological age in blood samples from deceased individuals.

Table 4.12: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* in blood samples from deceased individuals.

Table 4.13: MAD between predicted and chronological ages stratified by age group in the training set of 47 blood samples from deceased individuals.

Table 4.14: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot assay in 62 blood samples from deceased individuals.

Table 4.15: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*, selected by stepwise regression approach, in blood samples from deceased individuals.

Table 4.16: MAD between predicted and chronological ages stratified by age group in the training set of 59 blood samples from deceased individuals.

Table 4.17: Comparison of age-correlated values obtained in blood samples from deceased Portuguese individuals through Sanger and SNaPshot methodologies.

Table 4.18: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in the overall sample set of blood samples from living and deceased individuals.

Table 4.19: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, selected by stepwise regression approach, in the overall set of blood samples from living and deceased individuals.

Table 4.20: MAD between predicted and chronological ages stratified by age group in the overall training set of 141 blood samples from living and deceased individuals.

Table 4.21: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot in 121 blood samples from living and deceased individuals.

Table 4.22: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*, selected by stepwise regression approach, in the overall set of blood samples from living and deceased individuals.

Table 4.23: MAD between predicted and chronological ages stratified by age group in the overall training set of 115 blood samples from living and deceased individuals.

Table 4.24: Comparison of age-correlated values obtained in blood samples from living and deceased Portuguese individuals through Sanger and SNaPshot methodologies.

Table 4.25: Multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test the association between the DNAm levels and chronological age in blood samples from deceased individuals using the Sanger sequencing methodology.

B. DNA methylation age estimation in tooth samples

Table 4.26: Comparison of two regression lines between males and females in tooth samples using data obtained from Sanger sequencing.

Table 4.27: Simple linear regression statistics of the best age predictors in *FHL2*, *PDE4C* and *ELOVL2* genes to test the association between DNAm levels obtained by bisulfite sequencing and chronological age in tooth samples from living and deceased individuals.

Table 4.28: Statistical parameters obtained by simple linear regression for the *FHL2* CpG4, selected by stepwise regression approach in tooth samples.

Table 4.29: Comparison of two regression lines between males and females in tooth samples using data obtained from the SNaPshot methodology.

Table 4.30: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using SNaPshot assay in teeth from living and deceased individuals.

Table 4.31: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *ELOVL2* and *KLF14* selected by stepwise regression approach, in tooth samples.

Table 4.32: Comparison of age-correlated values obtained in tooth samples from living and deceased Portuguese individuals through Sanger and SNaPshot methodologies.

C. DNA methylation age estimation in fresh bone samples

Table 4.33: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in bone samples from autopsies.

Table 4.34: Statistical parameters obtained in a multiple regression model with the six CpGs in genes *ELOVL2*, *C1orf132* and *EDARADD* selected by stepwise regression approach, in bone samples from autopsies.

Table 4.35: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot assay in bones from autopsies.

Table 4.36: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *FHL2* and *KLF14* selected by stepwise regression approach, in bone samples from autopsies.

Table 4.37: Comparison of age-correlated values obtained in bones from autopsies through Sanger and SNaPshot methodologies.

Table 4.38: Comparison of two regression lines between males and females in bones from BDS using data obtained from Sanger sequencing.

Table 4.39: Simple and multiple linear regression statistics of the best age predictors in *FHL2*, *EDARADD* and *PDE4C* genes using bisulfite sequencing to test for association between the DNAm levels and chronological age in bones from BDS.

Table 4.40: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *FHL2* and *PDE4C* selected by stepwise regression approach, in bones from BDS.

Table 4.41: Simple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using SNaPshot in bones from BDS.

Table 4.42: Comparison of age-correlated values obtained in bones from BDS through Sanger and SNaPshot methodologies.

Table 4.43: *Postmortem* interval of the selected BDS with the same or similar chronological age.

Table 4.44: Average of the differences between DNAm levels of all the CpGs from each gene for the two individuals included in each pair A, B, C, D, E and F.

D. DNA methylation age estimation in buccal swabs

Table 4.45: Comparison of two regression lines between males and females in buccal swabs from living individuals using data obtained from Sanger sequencing.

Table 4.46: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2* gene to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in buccal swabs from living individuals.

Table 4.47: Statistical parameters obtained in a multiple regression model with the two CpGs in *ELOVL2* gene, selected by stepwise regression approach, in buccal swabs from living individuals.

Table 4.48: MAD between predicted and chronological ages stratified by age group in the training set of 23 buccal swabs from living individuals.

Table 4.49: Comparison of two regression lines between males and females in buccal swabs from living individuals using data obtained from the SNaPshot methodology.

Table 4.50: Simple linear regression statistics at the three CpGs of the *ELOVL2*, *KLF14* and *TRIM59* genes using SNaPshot assay in buccal swabs from living individuals.

Table 4.51: Statistical parameters obtained by simple linear regression for the *TRIM59* gene, selected by stepwise regression approach in buccal swabs.

Table 4.52: MAD between predicted and chronological ages stratified by age group in the training set of 35 buccal swabs from living individuals.

Table 4.53: Comparison of age-correlated values obtained in buccal swabs through Sanger and SNaPshot methodologies.

E. DNA methylation age estimation through multi-tissues

Table 4.54: Simple and multiple linear regression statistics of the age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels and chronological age in the group 1.

Table 4.55: Statistical parameters obtained in a multiple regression model with the nine CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, selected by stepwise regression approach, in blood, bone and tooth samples.

Table 4.56: Simple and multiple linear regression statistics of the age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between the DNAm levels and chronological age in the group 2.

Table 4.57: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2* and *C1orf132*, selected by stepwise regression approach, in blood, bone and tooth samples.

Table 4.58: Multi-tissue APM based in *ELOVL2* methylation information obtained using Sanger sequencing methodology.

Table 4.59: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *KLF14*, *TRIM59* and *C1orf132* genes using SNaPshot assay in the group 1.

Table 4.60: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *C1orf132* and *KLF14*, selected by stepwise regression approach, in blood, bone and tooth samples.

Table 4.61: Simple and multiple linear regression statistics at the three CpGs of the *ELOVL2*, *KLF14* and *TRIM59* genes using SNaPshot assay in the group 2.

Table 4.62: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *ELOVL2* and *KLF14*, selected by stepwise regression approach, in blood, bone, tooth and buccal swab samples.

F. Dry bone samples from *Coleção de Esqueletos Identificados do Século XXI - CEI/XXI*

Table 4.63: *Antemortem* data of selected skeletons from CEI/XXI.

Table 4.64: Design research made using bones from CEI/XXI.

Table 4.65: Quantification data obtained in dry bones from CEI/XXI with the standard protocol.

Table 4.66: Quantification data obtained in dry bones from CEI/XXI with the alternative Protocol 1 – EDTA.

Table 4.67: Methylation quantification of CpGs located at *ELOVL2* and *C1orf132* genes using the femur sample of the skeleton CEI/XXI 87.

Table 4.68: Quantification data obtained in dry bones from CEI/XXI with the alternative Protocol 2 – EDTA.

Chapter 5. General discussion

Table 5.1: Age-correlated values for the best CpG sites in each tissue type using Sanger sequencing.

Table 5.2: Age-correlated values obtained for the CpGs located at genes *ELOVL2*, *FHL2*, *KLF14*, *TRIM59* and *C1orf132* in each tissue type using the SNaPshot assay.

Table 5.3: Summary of the best tissue-specific APMs based on DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* captured by Sanger sequencing and SNaPshot methodologies.

Table 5.4: Summary of the multi-locus multi-tissue APMs based on DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* captured by Sanger sequencing and SNaPshot methodologies.

List of Figures

Chapter 1. General introduction

Figure 1.1: Chronological and biological ages from two different individuals (represented by blue and orange colors) who were born on the same day (individuals with the same chronological age). In early stage of development, individuals share the same biological age, which is consistent with the real age or chronological age. During their lifespan, they may have different predispositions to several conditions (as diseases) and may die at different periods. This will be reflected in their biological age. Adapted from Field *et al.* (2018).

Figure 1.2: DNAm. A) The addition of CH₃ group transforms unmethylated cytosine in C^m. B) Unmethylated and methylated CpG dinucleotides affect the gene expression. Methylated CpGs induce gene repression and unmethylated CpGs lead to gene transcription. Figure adapted from Gillespie *et al.* (2019).

Figure 1.3: Epigenetic drift and epigenetic clock. Figure adapted from Jones *et al.* (2015).

Figure 1.4: Approach proposed to the development of APMs based on DNAm levels. Figure adapted from Goel *et al.* (2017).

Chapter 2. Objectives

Chapter 3. Sample and design research

Figure 3.1: Blood samples from deceased individuals collected during autopsies in INMLCF.

Figure 3.2: Bones. A) Bone sample collected from one deceased individual during autopsy in INMLCF; B) Bone sample collected from BDS; C) Bone sample collected from BDS, the photo was captured after bone cut.

Figure 3.3: Dry bone sample from CEI/XXI, the photo was captured after bone cut.

Figure 3.4: Tooth samples in plastic recipient.

Figure 3.5: Buccal swabs from living individuals.

Figure 3.6: Bone fragments in vial before grinding.

Figure 3.7: A) Workstation II; B) Bone powder before the addition of prepared PrepFiler BTA™ lysis solution.

Page | XXII

Figure 3.8: A) Robot AutoMate Express™ Forensic DNA Extraction System from INMLCF (Coimbra, Delegação do Centro); B) Isolated DNA.

Figure 3.9: First step of DNA extraction from buccal swabs: resuspension of swabs in 1 ml of distilled water.

Figure 3.10: DNAm analysis based on bisulfite-PCR sequencing. After bisulfite conversion, in CpG sites (red circles) only the unmethylated cytosine is converted into U, while the 5mC or C^m remains as C. In non-CpG sites, the unmethylated cytosine is also transformed in U (red arrows). After PCR amplification all the U are displayed as T in the Sanger sequence of the sense strand. Adapted from GrŠković *et al.* (2013).

Figure 3.11: Sequencing data after bisulfite conversion and PCR amplification. Original DNA sequence reveals non-CpG sites and CpG sites. In chromatogram, red arrows represent converted cytosines at non-CpG sites and blue arrows show the CpG sites. Methylated cytosines at three CpG sites are represented by Y in the converted DNA sequence from the gene *ELOVL2*. G and A represent guanine and adenine, respectively.

Figure 3.12: Peaks height determination in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium). The example shows the evaluation of DNAm levels of CpGs located at *EDARADD* locus in a younger adult individual (28 years, female). In each CpG, the blue arrow (peak height C) represents methylated cytosines and the red arrow (peak height T) represents unmethylated cytosines. The value of the blue and the red arrows are putted in the formula $[C/C+T]$ and the percentage of DNAm level is calculated.

Figure 3.13: Examples of sequencing chromatograms for different methylation levels of the *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132* genes by bisulfite PCR-sequencing. Blue arrows show CpG sites from each gene. A) Chromatogram of the *ELOVL2* gene for a blood sample from a living individual (37 years, female); B) Chromatogram of the *EDARADD* gene for a blood sample from a living individual (22 years, female); C) Chromatogram of the *FHL2* gene for a blood sample from a living individual (2 years, male); D) Chromatogram of the *PDE4C* gene for a blood sample from a living individual (94 years, female); E) Chromatogram of *C1orf132* gene for a blood sample from a deceased individual (35 years, male).

Figure 3.14: Agarose gel electrophoresis with the amplification products obtained using multiplex methylation SNaPshot assay for five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes. The first band represents the products of amplification of the *ELOVL2* (187bp) and *FHL2* (191bp), the second the amplification product of *TRIM59* (148bp) and the third band corresponds to the amplification products of *C1orf132* (116bp) and *KLF14* (114bp). NTC, negative PCR control; L, DNA Ladder.

Figure 3.15: Methylation SNaPshot assay. Adapted from Quick Reference Card: ABI PRISM® SNaPshot™ Multiplex System (Applied Biosystems, 2010).

Figure 3.16: Electropherogram of the multiplex methylation SNaPshot assay for detection of DNAm levels in a blood sample from a female with 51 years old. Black (C) and blue (G) peaks represent methylation signal; red (T) and green (A) peaks represent non-methylation signal.

Chapter 4. Results and discussion

A. DNA methylation age estimation in blood samples

Figure 4.1: Predicted age *versus* chronological age using the four best markers *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 in blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.2: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from living individuals.

Figure 4.3: MAD from chronological age calculated for each age group in blood samples from living individuals. The MAD is printed on top of each respective age range.

Figure 4.4: Predicted age *versus* chronological age of the test set, 18 living individuals, using the final model developed for blood samples from living individuals with the markers *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2. MAD value and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.5: MAD from chronological age in the test set, 18 samples, using the developed APM for blood samples from living individuals. The MAD values are printed on the top of each age range. Age groups 1 and 2 were grouped together.

Figure 4.6: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay developed with the three CpGs located at *ELOVL2*, *FHL2* and *C1orf132*

genes in blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.7: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from living individuals.

Page | XXIV

Figure 4.8: MAD from chronological age calculated for each age group in blood samples from living individuals. The MAD is printed on top of each respective age range.

Figure 4.9: Predicted age *versus* chronological age using the five best markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *PDE4C* CpG2 and *C1orf132* CpG1 in blood samples from deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.10: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from deceased individuals.

Figure 4.11: MAD from chronological age calculated for each age group in blood samples from deceased individuals. The MAD is printed on top of each respective age range.

Figure 4.12: Predicted age *versus* chronological age of the test set, 19 deceased individuals, using the final best model developed for blood samples from deceased individuals including the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *C1orf132* CpG1 and *PDE4C* CpG2. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.13: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the four CpGs located at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in blood samples from deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.14: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from deceased individuals.

Figure 4.15: MAD from chronological age calculated for each age group in blood samples from deceased individuals. MAD increases with age to age group 3 (72-88 years old). The MAD is printed on top of each respective age range.

Figure 4.16: Predicted age *versus* chronological age using the four best markers *EDARADD* CpG3, *FHL2* CpG1, *ELOVL2* CpG6 and *PDE4C* CpG2 in blood samples from living and deceased individuals. MAD and Spearman correlation, r , are plotted on the chart.

Figure 4.17: Differences between chronological and predicted ages (years) plotted against chronological age (years) in the combined set of blood samples from living and deceased individuals.

Figure 4.18: MAD from chronological age calculated for each age group in blood samples from living and deceased individuals. The MAD is printed on top of each respective age range.

Figure 4.19: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the four CpGs located at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in blood samples from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.20: Differences between chronological and predicted ages (years) plotted against chronological age (years) in the combined set of blood samples from living and deceased individuals.

Figure 4.21: MAD from chronological age calculated for each age group in blood samples from living and deceased individuals. The MAD is printed on top of each respective age range.

Figure 4.22: Predicted age *versus* chronological age in a test sample set of blood samples from deceased individuals evaluated through Sanger sequencing methodology. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.23: Predicted age *versus* chronological age in a test sample set of blood samples from deceased individuals evaluated through SNaPshot methodology. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.24: Predicted age *versus* chronological age using the APM developed with the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3 and *PDE4C* CpG2 genes in the training set of deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.25: Predicted age *versus* chronological age in a test sample set of 53 blood samples from living individuals using the new model developed using blood from deceased individuals with the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3 and *PDE4C* CpG2 markers. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.26: Predicted age *versus* chronological age in a test sample set of blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

B. DNA methylation age estimation in tooth samples

Figure 4.27: Predicted age *versus* chronological age using best marker *FHL2* CpG4 in teeth from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.28: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the two CpGs located at *ELOVL2* and *KLF14* genes in teeth from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

C. DNA methylation age estimation in fresh bone samples

Figure 4.29: Predicted age *versus* chronological age using the six markers *ELOVL2* CpG5, *ELOVL2* CpG6, *ELOVL2* CpG7, *EDARADD* CpG3, *EDARADD* CpG4 and *C1orf132* CpG1 in bones from autopsies. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.30: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the two CpGs located at *FHL2* and *KLF14* genes in bones from autopsies. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.31: Predicted age *versus* chronological age using the two markers *FHL2* CpG2 and *PDE4C* CpG3 in bones from BDS. MAD and Spearman correlation coefficient, r , are plotted on the chart.

D. DNA methylation age estimation in buccal swabs

Figure 4.32: Predicted age *versus* chronological age using the two markers *ELOVL2* CpG1 and *ELOVL2* CpG4 in buccal swabs from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.33: Differences between chronological and predicted ages (years) plotted against chronological age (years) in buccal swabs from living individuals.

Figure 4.34: MAD from chronological age calculated for each age group in buccal swabs from living individuals. MAD increases with age. The MAD is printed on top of each respective age range.

Figure 4.35: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the CpG site located at *TRIM59* gene in buccal swabs from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Figure 4.36: Differences between chronological and predicted ages (years) plotted against chronological age (years) in buccal swabs from living individuals.

Figure 4.37: MAD from chronological age calculated for each age group in buccal swabs from living individuals. MAD increases with age. The MAD is printed on top of each respective age range.

E. DNA methylation age estimation through multi-tissues

Figure 4.38: Predicted age *versus* chronological age using the multi-locus multi-tissue APM developed for *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4) and tooth samples from deceased individuals (5).

Figure 4.39: Predicted age *versus* chronological age using the multi-tissue model developed for *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes including blood samples from deceased individuals (2), bone samples (3) and tooth samples from living (4) and deceased individuals (5).

Figure 4.40: Predicted age *versus* chronological age using the multi-tissue APM developed for *ELOVL2*, *C1orf132* and *KLF14* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4) and tooth samples from deceased individuals (5).

Figure 4.41: Predicted age *versus* chronological age using the multi-tissue APM developed for *ELOVL2*, *KLF14* and *TRIM59* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4), tooth samples from deceased individuals (5) and buccal swabs (6).

Figure 4.42: A) Incubation of bone powder of skeleton CEI/XXI 87 with Protocol 1 – EDTA. B) Bone powder of several skeletons form CEI/XXI (around 100-150 mg) before the incubation with EDTA.

Chapter 5. General discussion

Chapter I. General introduction

Chapter I. General introduction

1. Forensic anthropology and forensic genetics

Forensic sciences, as Forensic Anthropology (FA) and Forensic Genetics (FG), have experienced an adaptation to the constant progress of the forensic field along the years. Despite FA has arisen as a subfield of Biological Anthropology, in these days FA is an independent discipline of forensic sciences that applies and adapts some methods of Biological Anthropology in forensic cases (Cunha and Ferreira, 2021). FA focuses on the study of skeletons, bodies in advanced state of decomposition, bone fragments and burnt bones but, at the same time, have an important role in the examination of the living, specifically in age estimation of the living (Cunha and Ferreira, 2021). The two main aims of FA are identification of human remains and assist to the forensic pathologist in the discussion of the cause and circumstance of death (Pinheiro *et al.*, 2015; Cunha and Ferreira, 2021). However, nowadays, Forensic anthropologists have a larger range of action, contributing to:

i) differentiation between *antemortem*, *perimortem* and *postmortem* trauma (Cunha, 2019);

ii) differentiation between the different types of trauma (sharp trauma, blunt trauma or gunshot trauma) (Blau, 2016; Love and Wiersema, 2016);

iii) evaluation of the time since death (*Postmortem* interval) through an holistic approach combining other scientific sciences as entomology and botany (Cunha, 2019);

iv) having an important role in scenarios of disaster victim identification and human rights violations, working in multidisciplinary teams (de Boer *et al.* 2019; Ubelaker *et al.*, 2019a; Cunha and Ferreira, 2021).

Page | 4

Indeed, FA has changed through the years, being recognized as a robust and scientific discipline (Blue, 2018). Moreover, several advances in technology and mathematical approaches, as the use of imaging techniques (Garvin and Stock, 2016), and the development of powerful statistical programs, have improved several fields of FA (Ubelaker and Khosrowshahi, 2019; Navega and Cunha, 2020).

Despite having a crucial role in human identification through the establishment of biological profile, FA has been considered as a secondary identifier. The Interpol (International Criminal Police Organization) guidelines consider fingerprints, DNA analysis, and odontology as the primary identifiers. However, with the current use of FA in several forensic contexts, it has been defended by the board members of the Forensic Anthropology Society of Europe (FASE) that it can be used as a “means of personal identification, particularly in situations with limited availability of traditional identification methods (i.e. dactyloscopy, odontology, and molecular genetic analysis)” (de Boer *et al.*, 2020, pp. 1).

FG is a field of forensic sciences that focuses in “the application of genetics (in the sense of a science with the purpose of studying inherited characteristics for the analysis of inter- and intraspecific variations in populations) to the resolution of legal conflicts” (Carracedo, 2013, pp. 19). The growth of genetics in forensic contexts cannot be explained by the increase of the number of forensic DNA laboratories or by the superficial improvement of basic techniques and equipment; essentially, the increase was related with the creation of new methodologies and techniques that suppress pre-existing problems (Cabo, 2012). Technical advances and the concern with quality assurance have

led to the evolution of the discipline, being accepted as a robust and reliable forensic tool nowadays. DNA analysis can provide a relevant source of information for forensic purposes. For example, analysis of STR (short tandem repeats) is considered the gold standard procedure for human identification process (Marano and Fridman, 2019). DNA analysis through a comparative process between samples recovered from crime scenes with suspects and possibly victims leads to positive identification. Also, other relevant features can be assessed by DNA analysis helping in forensic investigations as externally visible characteristics of the individual (such as color of skin and hair), ancestry (assessed by biogeographic markers) (Kayser, 2015; Marano and Fridman, 2019; Samuel and Prainsack, 2019) and age estimation (assessed by epigenetic features) (Zolotareno *et al.*, 2019).

2. Age estimation in forensic contexts

Age estimation is a paramount issue in forensic science. Essentially, age estimation is required for identification of dead individuals; however, due to the large amount of recent humanitarian crises, it is increasingly essential to estimate age in living individuals (Schmeling *et al.*, 2011; Franklin *et al.*, 2015). The estimation of age can have different objectives: at the one hand, for dead bodies including human skeletonized remains, the main objective of age estimation is to assist in the identification through the establishment of the biological profile (sex estimates, ancestry inference, age-at-death estimation and stature estimated). Age-at-death, by itself, can lead to an exclusion. On the other hand, aging the living can be essential to solve judicial, criminal or civil problems, such as immigration cases (where the identity and age of individuals are unclear), cases of minor (regarding questions of adoption, imputability, pedo-

pornography), for determination of criminal responsibility, or even in civil cases of pensionable age (for old adults lacking documents) (Cunha *et al.*, 2009; Franklin *et al.*, 2015; Parson, 2018; Nuzzolese and Di Vella, 2019).

Page | 6

One of the big challenges for age estimation research is the establishment of “a standard method” that could be applied to several contexts. However, there is not one unique standardized method that can be applied for living and deceased individuals and/or to all age categories with the same accuracy (Cunha *et al.*, 2009; Baccino *et al.*, 2013). Therefore, the multidisciplinary approach to age estimation is a crucial issue.

For forensic anthropological research, based on published data from 2016 to 2019, age-at-death estimation is the most investigated issue (with a total of 201 articles) considering four forensic journals: Journal of Forensic Sciences, International Journal of Legal Medicine, Forensic Science International and Forensic Sciences Research, (Ubelaker *et al.*, 2020). In the past years, several areas have been focused in age estimation research, leading to the development of specific methods in distinct fields such as anthropology, odontology, chemistry, genetics and epigenetics, among others (Adserias-Garriga *et al.*, 2018; Adserias-Garriga, 2019a). Recently, it has been defended that the use of combined approaches based on several areas of research allowed for a more accurate age assessment. A study proposed by Shi *et al.* (2018) revealed that the combination of anthropological and epigenetic methods leads to an increase of age estimation accuracy. Likewise, the use of molecular approaches as the quantification of T-cell specific DNA rearrangements (sjTREC) combined with the evaluation of DNAm levels can lead to an improvement of age prediction accuracy, being mainly notorious in older age ranges (Cho *et al.*, 2017). Moreover, Becker *et al.* (2019) revealed that combining molecular approaches (racemization of D-aspartic acid combined with the

accumulation of pentosidine) in different tissues allow to obtain a high accuracy in age estimation.

Furthermore, it must be considered that any age estimation approach leads to estimate the biological age of an individual and not the chronological age. Chronological age refers to the time since birth, while biological age is related to alterations in the organism with lifespan, being associated with clinical conditions or age-related diseases (Horvath and Raj, 2018; Field *et al.*, 2018). Due to genetic, epigenetic and environmental influences (as lifestyle choices and diseases), biological age can differ substantially from chronological age (Field *et al.*, 2018, **Figure 1.1**). Hence, with the continuous growth of an individual or exposure to different environments, the difference between the biological and chronological ages increases (Baccino *et al.*, 2013).

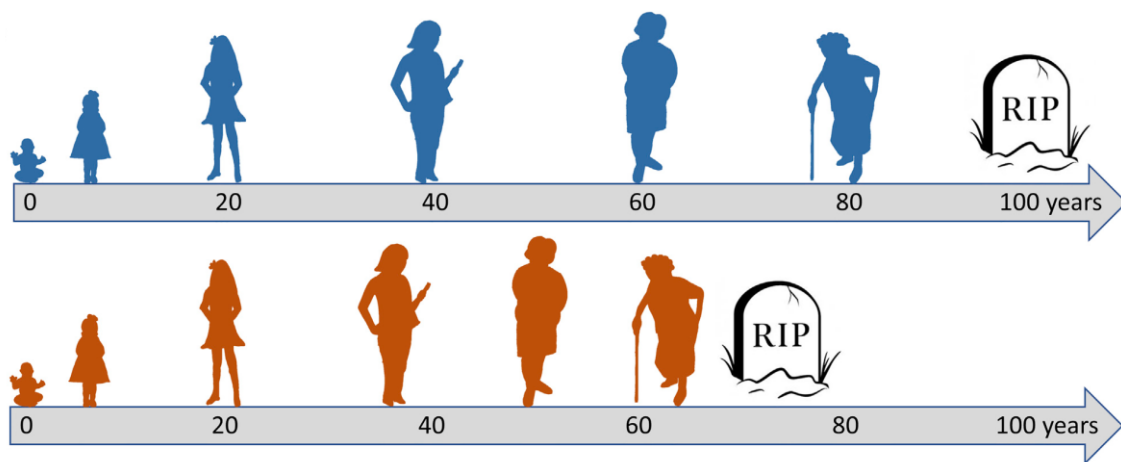


Figure 1.1: Chronological and biological ages from two different individuals (represented by blue and orange colors) who were born on the same day (individuals with the same chronological age). In early stage of development, individuals share the same biological age, which is consistent with the real age or chronological age. During their lifespan, they may have different predispositions to several conditions (as diseases) and may die at different periods. This will be reflected in their biological age. Adapted from Field *et al.* (2018).

Finally, it should be noted that when choosing a scientific method to estimate age, it is necessary to account for many aspects such as the cost of the method, the time spent

in obtaining results, whether the method is invasive or noninvasive or whether the obtained results are reliable for forensic purposes. Any method that involves DNA analysis turns out to be more expensive, but it also tends to be more reliable. However, it will be a destructive method for the forensic sample. Even more, if the sample is older or degraded, the fidelity of the results can be affected by the difficulty in obtaining suitable DNA for forensic analysis. Anthropological methods, based in skeletal or dental indicators, can overcome the sample destruction limitation for instance, but they are more subjective and may depend of the inter-observer error (Baccino *et al.*, 1999; Cunha *et al.*, 2009; Franklin, 2010; Nakhaeizadeh *et al.*, 2014).

2.1. Anthropological age estimation methods

Forensic anthropology has always played a key role in age estimation, namely age-at-death. However, in recent years, aging the living has become more and more relevant in forensic contexts as in cases related to imputability, refugees, illegal immigrants, and pensionable situations, which require the expertise of FA.

According to the recommendations of The Study Group on Forensic Age Diagnostics (Schmeling *et al.*, 2011) and current guidelines of INMLCF (Azevedo *et al.*, 2019), age estimation in living non-adults should be based on:

- i) physical or clinical evaluation;
- ii) evaluation of bone development (radiographic examination of the left hand and wrist; the examination of the degree of fusion of the cartilage in an x-ray on the hand-wrist area is the selected method in ages under 18 years;
- iii) examination of dental development (orthopantomography).

For age estimation in living adults, there are no recommended guidelines until now. Despite this, currently, an evaluation of hormone dosage after menopause (Cattaneo

et al., 2008) and dental approaches as pulp chamber methods (Cameriere *et al.*, 2007, 2012) should be done.

For age assessment in dead individuals, the macroscopic examination is the crucial step. Bodies in an advanced stage of decomposition require the evaluation of bones or teeth for the application of age estimation methodologies. Moreover, the potential applicability of each method depends on the type of sample and its state of preservation and whether dealing with adult human remains or not (Cunha *et al.*, 2009; Baccino *et al.*, 2013). Age estimation for non-adults and adults is based on different age indicators (Adserias-Garriga and Wilson-Taylor, 2019; Cunningham, 2019). Age estimation in non-adults focuses on stages of growth and development of the skeleton and stages of the development and eruption of teeth, while in adults age estimation is based on macroscopic degenerative changes of bones and teeth. Age estimation in non-adults is more reliable than in adults.

In non-adults, dental age is more accurate than skeletal age because it is not so affected by environmental factors as nutrition or social conditions (Cunha *et al.*, 2009; Elamin and Liversidge, 2013; Cunningham, 2019). For dental age estimation in subadults (0–20 years) it is recommended to select schemas as the AlQahtani atlas based on tooth development (AlQahtani *et al.*, 2010). Moreover, we should note that the development of the teeth is better than the eruption. Meanwhile, for age estimation, it is essential to cross the dental and skeletal ages. Skeletal age indicators for subadults as skull bones ossification, fontanelles, bone growth (length of long bone diaphysis), and epiphyseal appearance and fusion should be checked (Cunningham *et al.*, 2016). The transition to adulthood can be marked, among others, by the fusion of spheno-occipital synchondrosis. Moreover, the appearance of thoracic and lumbar vertebral rings, and the development and eruption of the third molar, when present, can help. In the transition phase (21–30

years), some skeletal age indicators are crucial as the examination of the medial end of the clavicle, which full ossification occurs until 30 years, and the epiphyses of the iliac crest (Risser method), which fusion occurs between 14 to 23 years (Schaefer *et al.*, 2009; Cunha *et al.*, 2009; Cunha, 2012; Cunningham *et al.*, 2016). These are among the last skeletal ossifications. With increasing age, skeletal age indicators are more unpredictable, being more difficult to make age estimation (Cunha, 2012).

For adults age estimation (young and mature adult: 30–55 years or elderly >55 years), methods based on macroscopic degenerative changes of some skeletal regions as pubic symphyses (Brooks and Suchey, 1990), sternal rib end (Işcan *et al.*, 1984, 1984a, 1985; Işcan and Loth, 1986), auricular surface (Lovejoy *et al.*, 1985; Buckberry and Chamberlain, 2002) and acetabular surface (San-Millán *et al.*, 2017) are usually used as conventional methods (Garvin and Passalacqua, 2012; Priya, 2017; Adserias-Garriga and Wilson-Taylor, 2019). As based essentially on the degenerative process of skeletal indicators, age estimation in adults remains one of the most difficult tasks for forensic anthropologists (Cunha *et al.*, 2009; Cunha, 2012; Priya, 2017; Adserias-Garriga and Wilson-Taylor, 2019). Age estimation of adults can also be performed with dental methods such as the case of Lamendin method based on root transparency (Lamendin, 1973, Lamendin *et al.* 1993) or deposition of secondary dentin (Cameriere *et al.*, 2007, 2012), among others (Singh and Singal, 2017; Odzhakov and Apostolov, 2019).

The combination of two or more methods can be a good alternative as it is the case of TSP (two step-procedure), which consists in the application of the Suchey-Brooks method in the first step, and if the morphology of the pubic symphysis shows older aspect, in the second step we applied the Lamendin method. The combination of both methods allows a more accurate age estimation (Cunha, 2012; Baccino *et al.*, 2013). This is due to the fact of each method works better between specif age ranges. While the Suchey-Brooks

method is more accurate between 20 and 40 years, the Lamendin method works better between 40 and 60 years, and the Iscan method was more advantageous in individuals older than 60 years (Cunha *et al.*, 2009). For age assessment in older age ranges, it is also recommended to observe secondary indicators such as acetabular degenerations, the translucency of cranial bones, degenerative pathologies, and osteoporosis (Teixeira and Cunha, 2021).

It is argued that combined anthropological indicators that reflect or capture more age-related aspects allow more accurate age estimations than the use of single methods (Baccino *et al.*, 1999; Martrille *et al.*, 2007; Priya, 2017). Despite that, we should note that combining methods based on different age indicators is also a difficult task. Each method was developed considering a specific reference sample with particular features (that can vary in population origin or ancestry, number of samples included, age of the individuals, among other factors). For instance, the Suchey-Brooks method is based on 1225 individuals (14-99 years old), of which only 273 are female with birth certificates and used in the method (Brooks and Suchey, 1990). Moreover, in regards to the age distribution of individuals, as essentially younger individuals (male and female) are included in the model, age estimation of older age categories can be affected. Consequently, the selection of the most suitable methods for anthropological age estimation in adults remains a challenge.

In conclusion, despite having several methods for forensic age estimation in living and dead people, it is more and more recommended the use of a multidisciplinary approach because there is not a unique age indicator that provides a perfect coincidence between biological age and chronological age, which is the age of interest in the forensic field (Adserias-Garriga and Wilson-Taylor, 2019).

2.2. Molecular and biochemical age estimation methods

Age estimations based on molecular methods, including shortening of telomeres, deletions of mitochondrial DNA (mtDNA) or signal joint T-cell receptor excision circles (sjTREC), or biochemical methods, including accumulation of D-aspartic acid and advanced glycation end products (AGEs), are being currently explored in forensic science community (Meissner and Ritz-Timme, 2010; Albert and Wright, 2015; Adserias-Garriga and Zapico 2018; Zapico *et al.*, 2019). These methods, based on the gradual alterations of biomolecules due the aging process, have been accepted as promising tools for age estimation. However, the lack of accuracy and other technical problems have limited the use of such methods in forensic casework (Meissner and Ritz-Timme, 2010; Albert and Wright, 2015; Zolotarenko *et al.*, 2019).

A brief review about some of these methods is detailed in **Table 1.1**.

Table 1.1: Molecular and biochemical methods used in age predictions.

Method	Relationship with age, applicability and advantages	Accuracy and forensic problems
Telomere length	Investigated in several samples as blood and skin (Butler <i>et al.</i> , 1998); blood (Tsuji <i>et al.</i> , 2002; Karlsson <i>et al.</i> , 2008; Ren <i>et al.</i> , 2009; Zubakov <i>et al.</i> , 2016) and tooth samples (Takasaki <i>et al.</i> , 2003; Márquez-Ruiz <i>et al.</i> , 2018, 2020), revealing a decrease of telomere length with the increase of age.	Small quantity of DNA or when it is degraded could affect the mean telomere length (Meissner and Ritz-Timme, 2010). High error of the estimate: Tsuji <i>et al.</i> (2002) observed in 60 blood samples and stains an age correlation coefficient of -0.832 and a standard error of 7.04 years; Takasaki <i>et al.</i> (2003), in dental pulp of 100 individuals, observed a correlation coefficient of -0.749 and a standard error of 7.52 years; Zubakov <i>et al.</i> (2016) observed a $R^2 = 0.141$ in 305 blood samples and an accuracy of 12.28 years; Márquez-Ruiz <i>et al.</i> (2018) investigating 91 teeth obtained a mean prediction error between predicted and chronological ages of 9.85 years and defended that this method should be used as a complementary approach to age estimations in forensic contexts. Moreover, they observed that the shortening can be affected by the type of teeth but not by the gender. More recently, Márquez-Ruiz <i>et al.</i> (2020) using 65 teeth obtained a significant correlation between age and telomere

		<p>length of -0.549 and observed a MAE (mean absolute error) of 6.89 years.</p> <p>The evaluation of telomere length seems to be influenced by several factors as cause of death and <i>postmortem</i> interval (Takasaki <i>et al.</i>, 2003), life choices and other environmental factors (Karlsson <i>et al.</i>, 2008). Moreover, it is associated with several diseases as obesity and cancer, being more advantageous in medicine than in forensics (Zolotarenko <i>et al.</i>, 2019).</p>
<p>Mitochondrial DNA (mtDNA) mutations</p>	<p>Alterations in mtDNA (duplications, deletions and point mutations) can occurred during aging. In forensic filed, mutations in mtDNA including deletions and point mutations can be used for age assessment (Zapico and Ubelaker, 2013). The mtDNA 4977 bp deletion, the most common studied deletion in age prediction, has been tested in many different tissues revealing an increase of deletions with the increase of age (Meissner and Ritz-Timme, 2010). The transition of adenine to guanine at the position 189 (A189G) is other mtDNA mutation that has been associated with age (Lacan <i>et al.</i>, 2011).</p>	<p>Mutations in mtDNA can vary between individuals and can be affected by several factors, such as UV radiation, diseases and other external factors (Zolotarenko <i>et al.</i>, 2019). It is not possible to estimate age with the high accuracy required for forensic contexts (Zolotarenko <i>et al.</i>, 2019).</p> <p>It is observed different frequencies of the mtDNA 4977 bp deletion in different tissues (Cortopassi <i>et al.</i>, 1992; Meissner <i>et al.</i>, 2008) and this deletion is not always present in all the aged individuals (Lee <i>et al.</i>, 1994; Diba <i>et al.</i>, 2015).</p>

<p>Aspartic acid Racemization (AAR)</p>	<p>Racemization (reaction that transforms L-amino acids to D-amino acids) occurs in all the amino acids, however aspartic acid racemization (AAR) is one of the fastest process. The D-aspartic acid content increases with increasing of age. It has been investigated in several tissues as dentin, cementum and bone, revealing high age correlation (Meissner and Ritz-Timme, 2010; Adserias-Garriga <i>et al.</i>, 2018). It is essentially good in dentin due to the accuracy, simplicity of the methodology and time required, compared with other dental tissues (Adserias-Garriga <i>et al.</i>, 2018).</p>	<p>Necessary the establishment of calibration curves for each tissue or protein analyzed (Meissner and Ritz-Timme, 2010). Often, destructive to bones and teeth.</p> <p>Not applicable to bodies exposed to high temperature (Ritz <i>et al.</i>, 1993).</p> <p>Ohtani and Yamamoto (2010) revealed an accuracy of ± 3 years in 5 cases reports. Wochna <i>et al.</i> (2018) reported an accuracy of 2.95-4.84 years in root dentin samples.</p>
<p>Signal joint T-cell receptor excision circles (sjTREC)</p>	<p>With the increase of age, the number of sjTREC decrease (Zubakov <i>et al.</i>, 2010; Ou <i>et al.</i>, 2011; Cho <i>et al.</i>, 2014; Zubakov <i>et al.</i>, 2016).</p> <p>Suitable for aged bloodstains (Zubakov <i>et al.</i>, 2010).</p> <p>Detected by PCR, available in forensic labs.</p>	<p>Low accuracy in several studies: Zubakov <i>et al.</i> (2010) using blood observed a $R^2 = 0.835$ and a standard error of the estimate of ± 8.9 years. Ou <i>et al.</i> (2011) using 245 blood samples from healthy individuals observed a correlation value of -0.818 and a standard error of ± 10.47 years. Cho <i>et al.</i> (2014) using 172 individuals observed and $r = -0.807$ and a standard error of 8.49 years. Zubakov <i>et al.</i> (2016) observed a little significant effect of sex in age predictions through the evaluation of sjTREC using 306 blood samples, revealing a $R^2 = 0.578$ and an accuracy of 9.39 years.</p>

3. Epigenetic age estimation

3.1 Evolution of epigenetics

Page | 16

The term “epigenetics” is not recent in scientific community and it has been redefined multiple times (Cavalli and Heard, 2019). Conrad Waddington was the first author to introduce it in the decade of 1940, as “the branch of biology that studies the casual interactions between genes and their products which bring the phenotype into being” (Waddington, 1942). After that, several authors have clarified the term “epigenetics” (Deans and Maggert, 2015). Nowadays, epigenetics represents a wide area of research that focuses on the mechanisms of gene control affecting gene expression without changing the DNA sequence. For example, Cavalli and Heard (2019) define epigenetics as “the study of molecules and mechanisms that can perpetuate alternative gene activity states in the context of the same DNA sequence” (pp. 489). Some epigenetic features are unique and specific for each individual, allowing the identification of several individual characteristics (Vidaki and Kayser, 2017, 2018; Williams, 2018).

The most investigated epigenetic mechanisms that are included in the epigenetic machinery are histone modifications, regulation by non-coding RNAs and DNA methylation (DNAm) (Roberti *et al.*, 2019). The study of alterations in these epigenetic features can bring relevant implications in clinical conditions and diseases, including cancer (Roberti *et al.*, 2019) and Alzheimer’s disease (Liu *et al.*, 2018). Moreover, some of these epigenetic features have been associated to the aging process (Lardenoije *et al.*, 2015).

3.2. DNA methylation (DNAm)

In eukaryotes, DNAm predominantly occurs in cytosine (C) residues, being an essential biochemical process characterized by the addition of a methyl group (CH_3) to the fifth carbon (5C) position of a C, resulting in 5mC (5-methylcytosine) or methylated cytosine (C^m) (**Figure 1.2**) (Schübeler, 2015; Jiang and Guo, 2020). Cytosine methylation is catalyzed by the action of three DNA methyltransferases (DNMT1, DNMT3A, DNMT3B), which establish and maintain DNAm in genome. DNMT1 keeps DNAm during cell division, and DNMT3A and DNMT3B are related with *de novo* methylation (Lyko, 2018).

In mammals, DNAm occurs commonly in dinucleotides CpG (5'-CpG-3' cytosine-phosphate-guanine). However, a smaller amount of DNAm can occur in non-CpG sites (CpH: H = A, C, T), being observed for instance in embryonic stem cells and neurons (He and Ecker, 2015).

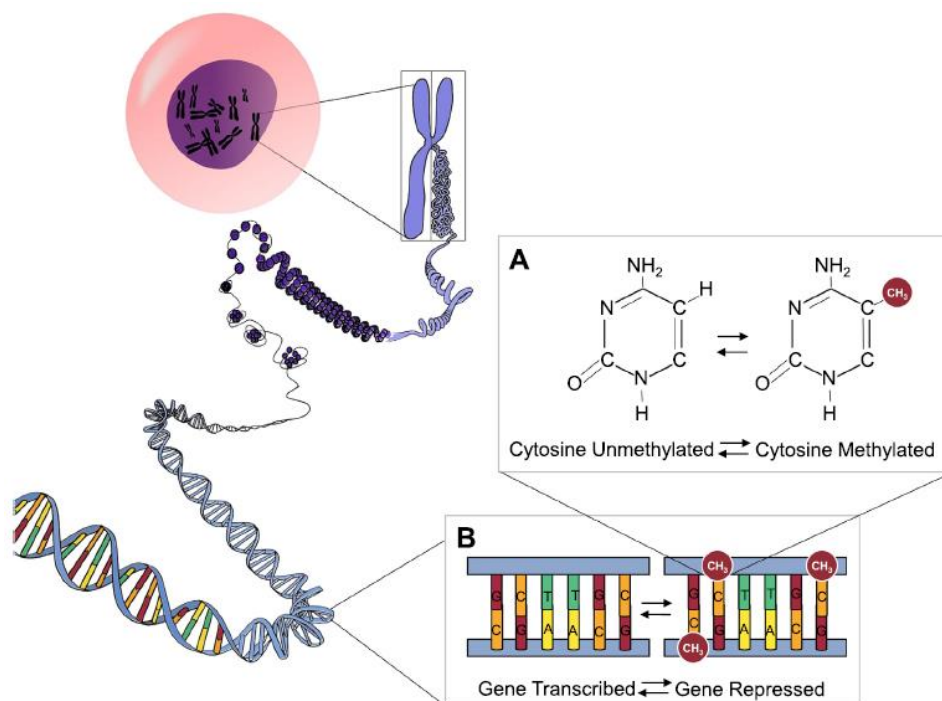


Figure 1.2: DNAm. A) The addition of CH_3 group transforms unmethylated cytosine in C^m . B) Unmethylated and methylated CpG dinucleotides affect the gene expression. Methylated CpGs induce gene repression and unmethylated CpGs lead to gene transcription. Figure adapted from Gillespie *et al.* (2019).

The human genome contains about 28 million of CpG dinucleotides, not uniformly distributed along the genome, which may be methylated or non-methylated. Most of the genome are methylated, however 20-40% of CpG sites are non-methylated (Deaton and Bird, 2011; Schübeler, 2015; Ciccarone *et al.*, 2018; Field *et al.*, 2018). Frequently, unmethylated CpG dinucleotides are located at clusters with high frequency of CpGs (CpG-rich regions) defined as CpG islands (CGIs) (for reviews, see Vidaki *et al.*, 2013; Kader and Ghai, 2015). Originally defined by Gardiner-Garden and Frommer in 1987, CGIs are considered as ≥ 200 bp long, having $\geq 50\%$ of GC content and $\text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}} \geq 0.60$ (Sarda and Hannenhelli, 2018). CGIs can be located in promoter regions, intragenic or intergenic regions (Illingworth *et al.*, 2010). About 60% of the CGIs are present in gene promoters and are frequently unmethylated being associated to gene expression of many genes (Larsen *et al.*, 1992; Freire-Aradas *et al.*, 2017). The remaining 40% are located at “orphan CpG islands”, along the intergenic and intragenic regions (Illingworth *et al.*, 2010; Zampieri *et al.*, 2015; Freire-Aradas *et al.*, 2017; Sarda and Hannenhelli, 2018; Ciccarone *et al.*, 2018). Some orphan CGIs (methylated or non-methylated, despite during the development of an organism are essentially methylated) may probably have a role as alternative promoters, being associated to tissue-specific gene expression (Illingworth *et al.*, 2010; Sarda and Hannenhelli, 2018).

Occurring commonly in CGIs, DNAm has been currently associated with gene expression and gene silencing. Through the influence of gene transcription, DNAm patterns are relevant for several process in organism such as development and differentiation, imprinting and X-chromosome inactivation, among others (Deaton and Bird, 2011; Moore *et al.*, 2013; Freire-Aradas *et al.*, 2017; Sarda and Hannenhelli, 2018). Moreover, environmental factors, as chemical exposures, nutritional status and lifestyle

choices (as tobacco consumption) can influence DNAm levels across the genome (Kader and Ghai, 2015; Martin and Fry, 2018; Prince *et al.*, 2019).

Changes in DNAm levels can be associated to many diseases or clinical conditions, including obesity and diabetes (Loh *et al.*, 2019) or cancer (Conway *et al.*, 2017). In forensic field, the evaluation of changes in DNAm levels has been explored in many investigations, including identification of biological fluids/tissues, sex determination, differentiation between monozygotic twins and ancestry determination (GrŠković *et al.*, 2013; Vidaki *et al.*, 2013; Kader and Ghai, 2015; Vidaki and Kayser, 2017, 2018; Williams, 2018). In particular, recent years have shown that DNAm is the epigenetic modification with most promising potential in developing methods for forensic age prediction (Horvath, 2013; Weidner *et al.*, 2014; Huang *et al.*, 2015; Kader and Ghai, 2015; Lee *et al.*, 2016; Zubakov *et al.*, 2016; Jylhävä *et al.*, 2017; Freire-Aradas *et al.*, 2017; Parson, 2018; Vidaki and Kayser, 2018; Williams, 2018; Zolotarenko *et al.*, 2019).

3.2.1. DNAm changes with age

Age-related changes in DNAm levels occur during life since the conception. With aging, there is a change in DNAm patterns in the genome: CpGs islands often associated with gene promoters (frequently unmethylated) acquire DNAm and the remaining CpGs across most of the genome (with high DNAm) lose methylation, reflecting a global genome hypomethylation (Heyn *et al.*, 2012; Florath *et al.*, 2014; Weidner *et al.*, 2014; Zampieri *et al.*, 2015; Jones *et al.*, 2015; Xiao *et al.*, 2016).

The basis of changes in DNAm with aging may be related with two opposing phenomena: epigenetic drift, caused by stochastic factors, and epigenetic clock, caused by nonstochastic events (Jones *et al.*, 2015) (**Figure 1.3**). The epigenetic clock is a directional phenomenon in which DNAm levels regularly change with age. It refers to

consistent alterations in DNAm levels of specific sites of the human genome across different and not related individuals or tissue types (Jones *et al.*, 2015; Pal and Tyler, 2016). Consequently, these CpG sites associated with epigenetic clocks can be used to predict chronological age (Jones *et al.*, 2015; Horvath and Raj, 2018). In contrast, the concept of epigenetic drift is related with environmental and stochastic factors that change DNAm levels and lead to a DNAm divergence across individuals (Fraga *et al.*, 2005; Jones *et al.*, 2015; Tan *et al.*, 2016). The increase of DNAm divergence between monozygotic twins with aging has been explained by the epigenetic drift contribution (Fraga *et al.*, 2005). The basis of the epigenetic drift is not yet clear however, these associated DNAm changes can be the result of individual lifestyles (as smoking habits, alcohol consumption or physical activity) (Freire-Aradas *et al.*, 2017).

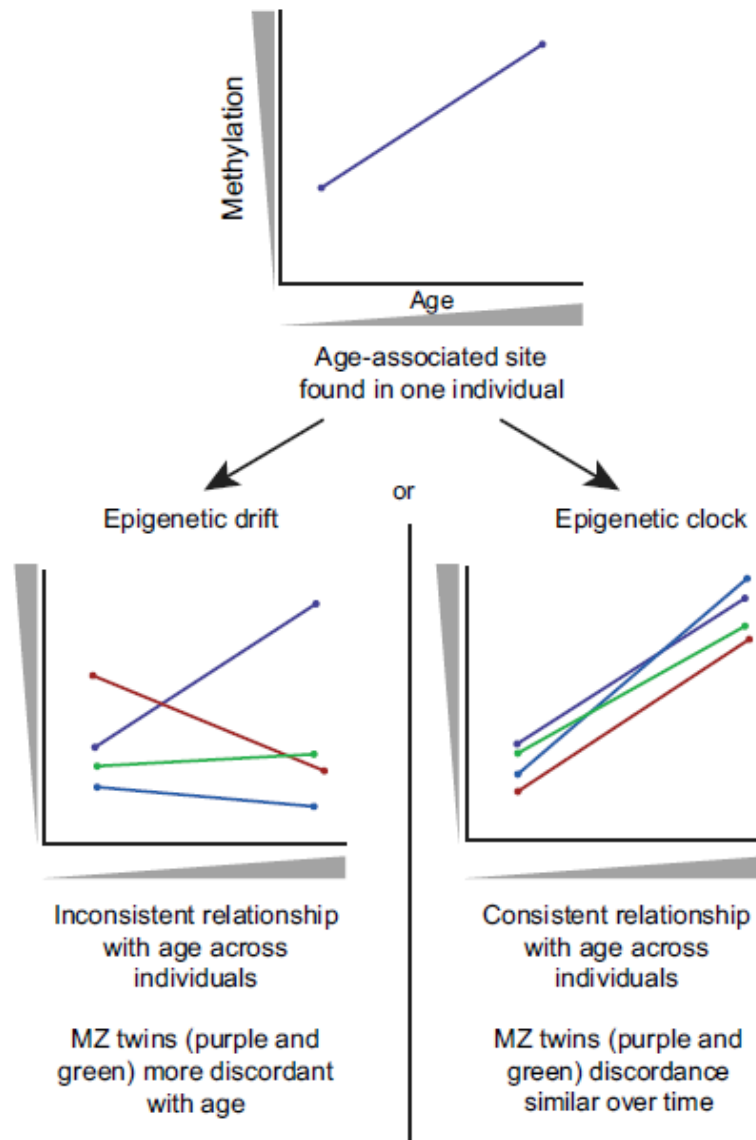


Figure 1.3: Epigenetic drift and epigenetic clock. Figure adapted from Jones *et al.* (2015).

Changes in the DNAm levels (due to the epigenetic drift and the epigenetic clock contributions) can be used to estimate the age of an individual. This estimated age through the measurement of DNAm changes is known as epigenetic age or DNAm age (Horvath and Raj, 2018).

In recent years, age estimation based on DNAm has been widely investigated for forensic purposes. It has been observed that several CpGs located at different genes, such as *ELOVL2*, *FHL2*, *NPTX2*, *EDARADD*, *ASPA*, *GRIA2*, *ITGA2B*, *PDE4C*, *PENK*, *CCDC102B*, *C1orf132*, *TRIM59* and *KLF14*, are strongly associated with chronological

age (Freire-Aradas *et al.*, 2017; Goel *et al.*, 2017; Zolotareno *et al.*, 2019). Consequently, a number of age prediction models (APMs) based on methylation information of these sites have been developed allowing to accurate age predictions. These APMs capture chronological age, but are also the reflection of the biological age (Chen *et al.*, 2016; Zhang *et al.*, 2017a; Field *et al.*, 2018; Horvath and Raj, 2018; Nwanaji-Enwerem *et al.*, 2018). Truly, differences between chronological and biological ages are influenced by age-related conditions as frailty (Breitling *et al.*, 2016) and age-related diseases, such as cancer (Horvath, 2013; Ambatipudi *et al.*, 2017; Dugué *et al.*, 2018) and Parkinson's disease (Horvath and Ritz, 2015).

3.2.2. Approaches used for the development of APMs based on DNAm

In general, most studies focusing on DNAm age take similar approaches to develop APMs, based on the following steps (Goel *et al.*, 2017): i) the evaluation of DNAm levels of a number of CpGs from specific tissue types; ii) the estimation of correlation coefficient values between each CpG site and chronological age, allowing the selection of the highly age-correlated CpGs from the investigated tissues; and iii) the use of multivariate regression analyses through different statistical approaches, to build the most accurate APMs (**Figure 1.4**).

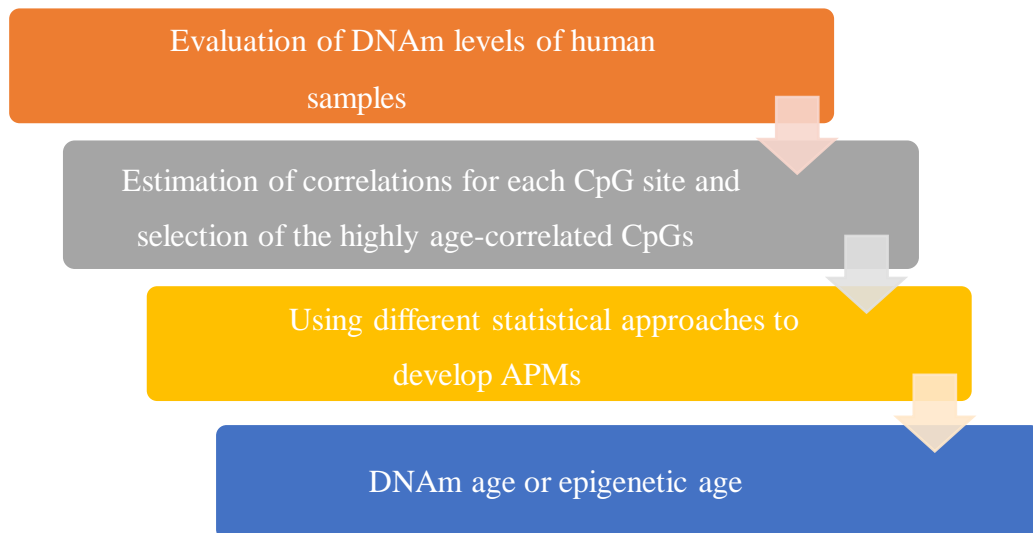


Figure 1.4: Approach proposed to the development of APMs based on DNAm levels. Figure adapted from Goel *et al.* (2017).

In recent years, several APMs have been developed differing in CpGs they contain, tissue types in which they were developed, statistical analyses or DNAm technologies (for review Goel *et al.*, 2017; Freire-Aradas *et al.*, 2017, 2020; Horvath and Raj, 2018; Liu *et al.*, 2019; Bergsma and Rogaeva, 2020).

From the several technologies addressing DNAm levels in the last decade, pyrosequencing was mainly used (Weidner *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Eipel *et al.*, 2016; Park *et al.*, 2016; Cho *et al.*, 2017; Thong *et al.*, 2017; Spólnicka *et al.*, 2017; Daunay *et al.*, 2019; Pfeifer *et al.*, 2020). Meanwhile, other methodologies also enabled DNAm profiling, including EpiTyper mass spectrometry (Garagnani *et al.*, 2012; Xu *et al.*, 2015; Giuliani *et al.*, 2016; Freire-Aradas *et al.*, 2016, 2018), massively parallel sequencing (MPS) (Vidaki *et al.*, 2017; Naue *et al.*, 2017, 2018; Aliferi *et al.*, 2018), SNaPshot (Lee *et al.*, 2015; Hong *et al.*, 2017, 2019; Jung *et al.*, 2019), methylation-sensitive high resolution melting (MS-HRM) (Hamano *et al.*, 2016, 2017) and microarray hybridization technology (Bocklandt *et al.*, 2011; Hannum *et al.*, 2013; Horvath, 2013; Florath *et al.*, 2014). Most of these methods employ

sodium bisulfite conversion, which mediates the deamination of unmethylated cytosine bases to uracil and leaves 5-methylcytosine. The technique of bisulfite sequencing is considered to be the “*gold standard*” method for DNAm analysis being a qualitative, quantifiable and effective method (Li and Tollefsbol, 2011; Patterson *et al.*, 2011). Very often, for forensics analysis it is accepted as the method of election (Vidaki *et al.*, 2013).

Most developed APMs are based on a limited number of CpGs associated with age, widely distributed throughout the genome, mainly identified and selected from DNAm array technology. These APMs have been developed based in blood samples (mainly from living individuals) (Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015a, 2015b; Huang *et al.*, 2015; Park *et al.*, 2016; Hamano *et al.*, 2016; Cho *et al.*, 2017; Naue *et al.*, 2017, 2018; Spólnicka *et al.*, 2017; Thong *et al.*, 2017; Alghanim *et al.*, 2017; Jung *et al.*, 2019; Daunay *et al.*, 2019; Pfeifer *et al.*, 2020), but other tissues and body fluids have also been explored, including saliva (Alghanim *et al.*, 2017; Hong *et al.*, 2017, 2019; Jung *et al.*, 2019), buccal swabs (Bekaert *et al.*, 2015b; Eipel *et al.*, 2016; Naue *et al.*, 2018; Jung *et al.*, 2019; Pfeifer *et al.*, 2020), semen (Lee *et al.*, 2015, 2018), teeth (Bekaert *et al.*, 2015a; Giuliani *et al.*, 2016; Márquez-Ruiz *et al.*, 2020), and bone (Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020). Moreover, few studies have investigated the stability of DNAm levels in bloodstains (Zbieć-Piekarska *et al.*, 2015a; Huang *et al.*, 2015; Thong *et al.*, 2017).

Nevertheless, the selection of universal age markers can be changeling for the development of accurate APMs due to the inter-tissue variability of DNAm patterns or tissue specificity of age-related DNAm changes (Freire-Aradas *et al.*, 2017; Jung *et al.*, 2017; Naue *et al.*, 2018; Sliker *et al.*, 2018; Jung *et al.*, 2019). The proposed age-associated CpGs in each tissue cannot be applied for age predictions in other tissue types, without being previously tested in those tissues. But at the same time, some markers can

be promising for age predictions across different tissue types, and DNAm changes of several age-dependent markers have been used in recent years to develop multi-tissue APMs (Horvath, 2013; Alsaleh *et al.*, 2017; Jung *et al.*, 2019). In any case, the identification of universal age biomarkers that could retain high correlation values with chronological age across different tissues, with similar accuracy in APMs, not being influenced by other conditions, could be a challenge task (Koch and Wagner, 2011; Horvath, 2013).

3.2.3. A brief review on DNAm age prediction models (APMs)

In last years, studies of DNAm analysis for forensic age prediction have grown in number, using a range of partially overlapping genes, different tissues and different technologies. A detailed review of the literature was carried out encompassing some of the most relevant scientific works to date. An overview of some of these developed APMs is presented in **Table 1.2**.

Table 1.2: Review of studies based on DNAm age estimation.

Study	Tissues	Age range	CpGs or genes	Technique	Accuracy of APM (training set)
Bocklandt <i>et al.</i> (2011)	60 saliva	Aged 18-70 years	<i>NPTX2, EDARADD, TOM1L1</i>	Illumina Infinium 27K	5.2 years (2 CpGs)
Hannum <i>et al.</i> (2013)	656 blood (training: 482; validation: 174)	Aged 19-101 years	71 markers located at several genes, including <i>ELOVL2, KLF14, TRIM59, Clorf132</i> , among others	Illumina Infinium 450K	3.88 years
Horvath (2013)	about 8000 samples, several tissues and cell types	Aged 0-101 years	353 CpGs Multi-tissue	Illumina Infinium 27K; Illumina Infinium 450K	2.9 years
Weidner <i>et al.</i> (2014)	151 blood (training: 82; validation: 69)	Aged 20-75 years	<i>ITGA2B, ASPA, PDE4C</i>	Pyrosequencing	5.4 years
Zbieć-Piekarska <i>et al.</i> (2015a)	303 blood	Aged 2-75 years	7 CpGs <i>ELOVL2</i>	Pyrosequencing	5.03 years (2 CpGs)
Zbieć-Piekarska <i>et al.</i> (2015b)	300 blood	Aged 2-75 years	<i>ELOVL2, Clorf132, TRIM59, KLF14, FHL2</i>	Pyrosequencing	3.4 years
Huang <i>et al.</i> (2015)	89 blood	Aged 9-75 years	<i>ASPA, ITGA2B, NPTX2</i>	Pyrosequencing	7.87 years
Lee <i>et al.</i> (2015)	68 semen	Aged 20-73 years	cg06304190 (<i>TTC7B</i>), cg12837463, cg06979108 (<i>NOX4</i>)	SNaPshot	4.7 years

Bekaert et al. (2015a)	206 blood (living: 37; deceased: 169)	Aged 0-91 years	4 CpGs at <i>ASPA</i> , <i>ELOVL2</i> , <i>PDE4C</i> , <i>EDARADD</i>	Pyrosequencing	3.75 years
	29 dentin (living)	Aged 19-70 years	7 CpGs at <i>PDE4C</i> , <i>ELOVL2</i> , <i>EDARADD</i>		4.86 years
Bekaert et al. (2015b)	50 buccal swabs	Aged 0-73 years	<i>ASPA</i> , <i>ELOVL2</i> , <i>PDE4C</i> , <i>EDARADD</i>	Pyrosequencing	3.32 years
Giuliani et al. (2016)	21 teeth (cementum, dentin and dental pulp)	Aged 17-77 years	<i>ELOVL2</i> , <i>FHL2</i> , <i>PENK</i>	Maldi-Tof mass spectrometry	2.25 years (pulp) 2.45 years (cementum) 7.07 years (dentin)
Eipel et al. (2016)	55 buccal swabs	Aged 1-85 years	cg17861230 (<i>PDE4C</i>), cg02228185 (<i>ASPA</i>), cg25809905 (<i>ITGA2B</i>)	Pyrosequencing	14.6 years
			Applied the Weidner model: “3-CpG-blood-model”		4.3 years
			“3-CpG-swab-model” “1-CpG-swab-model”		5.2 years
Zubakov et al. (2016)	216 blood	Aged 4-82 years	43 CpGs, including CpGs from <i>FHL2</i> , <i>ELOVL2</i>	EpiTYPER	4.23 years (43 CpGs) 5.09 years (8 CpGs)

Freire-Aradas <i>et al.</i> (2016)	725 blood	Aged 18-104 years	177 CpG sites	Illumina Infinium 450K, EpiTYPER	3.07 years (7 genes)
Hamano <i>et al.</i> (2016)	74 blood (living, 22; deceased, 52)	Aged 0-95 years	<i>ELOVL2, FHL2</i>	MS-HRM	7.44 years
Cho <i>et al.</i> (2017)	100 blood	Aged 20-74 years	32 CpGs located at <i>ELOVL2, FHL2, C1orf132, TRIM59, KLF14</i> Applied the predictive equation by Zbieć-Piekarska <i>et al.</i> (2015b) APM with best-selected sites in Koreans: <i>ELOVL2</i> (Chr6:11044628), <i>C1orf132</i> (Chr1:207823681), <i>TRIM59</i> (Chr3:160450189), <i>KLF14</i> (Chr7:130734355), <i>FHL2</i> (Chr2:105399282)	Pyrosequencing	4.18 years 3.34 years
Thong <i>et al.</i> (2017)	145 blood	Aged 3-80 years	<i>ELOVL2, FHL2, C1orf132, TRIM59, KLF14</i> APM with <i>ELOVL2</i> (Chr6:11044642), <i>TRIM59</i> (Chr3:160450189), <i>KLF14</i> (Chr7:130734357)	Pyrosequencing	3.3 years

Hong <i>et al.</i> (2017)	226 saliva (training: 113)	Aged 18-65 years	APM with 7 CpGs: cg18384097 (<i>PTPN7</i>), cg00481951 (<i>SST</i>), cg19671120 (<i>CNGA3</i>), cg14361627 (<i>KLF14</i>), cg08928145(<i>TSSK6</i>), cg12757011 (<i>TBR1</i>), cg07547549 (<i>SLC12A5</i>)	SNaPshot	3.13 years
			APM with 6 CpGs located at <i>SST</i> , <i>CNGA3</i> , <i>KLF14</i> , <i>TSSK6</i> , <i>TBR1</i> and <i>SLC12A5</i> (without cg18384097 located at <i>PTPN7</i>)		4.1 years
Lee <i>et al.</i> (2018)	12 semen	Aged 24-57 years	cg06304190 (<i>TTC7B</i>), cg12837463 and cg06979108 (<i>NOX4</i>); validation of the model by Lee <i>et al.</i> (2015)	SNaPshot	4.8 years
	19 forensic samples	Aged 17-48 years			5.2 years
Horvath <i>et al.</i> (2018)	Several tissues, including buccal swabs, saliva, blood, endothelial cells, skin among others	Aged 0-112 years	391 CpGs “skin & blood clock”	Illumina Infinium 450K and Infinium EPIC (or 850K)	2.9 years (blood) 2.5 years and 3 years (saliva)
Freire-Aradas <i>et al.</i> (2018)	180 blood	Aged 2-18 years	CpGs at <i>KCNAB3</i> ; <i>EDARADD</i> ; <i>ELOVL2</i> ; <i>CCDC102B</i> ; <i>MIR29B2CHG</i> genes and position CR_23_CpG3 (no gene)	EpiTYPER	0.94 years

Naue et al. (2018)	144 samples from 29 deceased individuals (bone, blood....)	Aged 0-87 years	13 blood age-correlated loci used in a previous study (Naue et al., 2017): <i>DDO</i> , <i>ELOVL2</i> , <i>F5</i> , <i>GRM2</i> , <i>HOXC4</i> , <i>KLF14</i> , <i>LDB2</i> , <i>MEIS1-AS3</i> , <i>NKIRAS2</i> , <i>RPA2</i> , <i>SAMD10</i> , <i>TRIM59</i> , <i>ZYG11A</i> .	MPS	-
Shi et al. (2018)	124 blood (48 blood by epigenetic approach)	Aged 6-15 years	Dental age and skeletal age + DNAm age (combined approach) Total of 485.577 CpG sites investigated; selection of CpGs located at <i>DDO</i> , <i>PRPH2</i> , <i>DHX8</i> , <i>ITGA2B</i> and at one unknown gene with the Illumina ID number of 22398226	Illumina Infinium 450K	0.47 years (boys) 0.33 years (girls)
Jung et al. (2019)	448 samples from blood, saliva and buccal swabs (training: 300, validation: 148)	Aged 18-74 years	<i>ELOVL2</i> , <i>FHL2</i> , <i>KLF14</i> , <i>C1orf132</i> , <i>TRIM59</i>	SNaPshot	3.55 years
Gopalan et al. (2019)	32 bones (preserved: 19, forensic: 13) 133 bones	Aged 18-97 years Aged 49-112 years	Forensic samples excluded. "37 bone clock CpGs" with CpGs located, among other genes, at <i>TRIM59</i> , <i>ELOVL2</i> and <i>KLF14</i>	Illumina Infinium 450K and Illumina EPIC	7.1 years (preserved) 4.9 years (all samples)

Márquez-Ruiz et al. (2020)	65 teeth	Aged 15-85 years	CpGs located at <i>ELOVL2</i> , <i>ASPA</i> and <i>PDE4C</i>	Pyrosequencing	5.08 years (<i>ELOVL2</i> , <i>PDE4C</i>)
Lee et al. (2020)	66 bones	Aged 31-96 years	<i>ELOVL2</i> , <i>FHL2</i> , <i>TRIM59</i> , <i>KLF14</i> , <i>C1orf132</i> genes proposed by Jung et al. (2019)	SNaPshot	–
			Applied the “skin & blood clock”	Infinium EPIC array	6.4 years
Pfeifer et al. (2020)	149 buccal swabs (living)	Aged 13 days-95 years	Applied the predictive equation by Bekaert et al. (2015a)	Pyrosequencing	9.84 years
			New model (<i>ASPA</i> , <i>ELOVL2</i> , <i>PDE4C</i> , <i>EDARADD</i>)		5.55 years
			Applied the predictive equation by Bekaert et al. (2015b)		8.68 years
			New model (<i>ASPA</i> , <i>ELOVL2</i> , <i>PDE4C</i> and <i>EDARADD</i>)		4.65 years

To the best of our knowledge, the first study of DNAm age estimation was performed by Bocklandt *et al.* (2011), testing DNAm levels of *NPTX2*, *EDARADD* and *TOM1L1* genes in 60 saliva samples (aged 18-70 years old) using the Illumina HumanMethylation27 microarray. An APM was built with only two sites, explaining 73% of the variation in age, with an average accuracy of 5.2 years. Later, Garagnani *et al.* (2012), using 501 blood samples (aged 9-99 years old), tested the DNAm of *ELOVL2*, *FHL2* and *PENK* loci using the Sequenom's EpiTYPER assay, showing the highest age correlation value 0.92 for *ELOVL2* gene. In 2013, Hannum *et al.* (2013) examined DNAm levels of more than 450.000 CpGs, captured from blood of 656 individuals (aged 19-101 years old), using the Illumina Infinium HumanMethylation450 BeadChip. Their final APM comprising 71 markers located at several genes, including *ELOVL2*, *KLF14*, *TRIM59*, and *C1orf132*, revealed a correlation value between predicted and chronological ages of 0.963 and high accuracy, with a root mean squared error (RMSE) of 3.88 years. Moreover, Hannum *et al.* (2013) evaluated the DNAm differences between different tissues (breast, kidney, lung, and skin samples) and built specific APMs for each tissue, observing that CpGs selected in each tissue were different between specific models. Meanwhile, all the models included CpGs from the *ELOVL2* gene. Thus, this gene showed to be a stable age marker displaying a similar and high degree of age-dependence across the different tissues. Also in this year, Horvath (2013), using the Illumina's Beadchip array, developed the first multi-tissue APM using about 8000 samples. The model with 353 CpGs showed a high correlation value between predicted and chronological ages of 0.97, and a prediction accuracy of 2.9 years in the training set. In 2015, Horvath *et al.* (2015) tested this model in 48 trabecular bone samples (aged 49-104 years old), obtaining a correlation value of 0.88 between predicted and chronological ages.

In 2014, Weidner *et al.* (2014) evaluated DNAm levels of only three CpGs from the *ITGA2B*, *ASPA* and *PDE4C* genes in 151 blood samples (training set: 82 samples and test set: 69 samples), using the bisulfite pyrosequencing methodology. A high accurate APM was developed with a Mean Absolute Deviation (MAD) between predicted and chronological ages of 5.4 years in the training set.

One year later, Zbieć-Piekarska *et al.* (2015a) studied seven CpGs from the *ELOVL2* locus, also by pyrosequencing using 303 blood samples (aged 2-75 years old). The developed final model with the two best CpGs, explained 85.9% of the variation in age, revealing a MAD from chronological age of 5.03 years. Moreover, investigating DNAm levels in 45 bloodstains, Zbieć-Piekarska and collaborators have shown that the DNAm of *ELOVL2* did not change after one-month storage as bloodstains. This first study was updated later with the analysis of five genes (*ELOVL2*, *Clorf132*, *TRIM59*, *KLF14* and *FHL2*) using a training set of 300 blood samples of Polish individuals (aged 2-75 years old) (Zbieć-Piekarska *et al.*, 2015b). The developed APM revealed a high correlation value between DNAm levels and age ($R = 0.971$) with a MAD from chronological age of 3.4 years. The authors developed an *age prediction calculator* available *online* (www.agecalculator.ies.krakow.pl). Another contribution in that year was made by Huang *et al.* (2015) evaluating DNAm levels from *ASPA*, *ITGA2B* and *NPTX2* genes in 89 blood samples (aged 9-75 years old) and 20 bloodstains from Chinese people using pyrosequencing, obtaining a MAD from chronological age of 7.87 years. Moreover, it has been shown a stable prediction in old bloodstains for 4-month storage (Huang *et al.*, 2015). Therefore, both reports of Huang *et al.* (2015) and Zbieć-Piekarska *et al.* (2015a) corroborate the DNAm stability in bloodstains.

The first study comprising blood samples from deceased individuals was made in 2015 by Bekaert *et al.* (2015a). Using pyrosequencing in 206 blood samples from living

and deceased individuals (aged 0-91 years old), they evaluated DNAm levels of *ASPA*, *ELOVL2*, *PDE4C* and *EDARADD* genes obtaining a predictive model accuracy of 3.75 years. They suggested similar distributions of DNAm changes when comparing blood samples from living and deceased individuals. Also, Bekaert *et al.* (2015a) evaluating DNAm of the same markers in 29 dentin samples from living individuals (19-70 years old), reported a model with seven CpGs located at *PDE4C*, *ELOVL2* and *EDARADD* genes, explaining 74% of the variation in age (corrected R^2 of 0.74), and a MAD of 4.86 years. This study is also the first report made in tooth samples. Furthermore, the same group evaluated DNAm of the *ASPA*, *ELOVL2*, *PDE4C* and *EDARADD* genes in 50 buccal swabs from living individuals (aged 0-73 years old), obtaining a MAD of 3.32 years (Bekaert *et al.* 2015b). In the same year, Lee and collaborators reported a semen-based APM developed in 68 samples with cg06304190 (*TTC7B*), cg12837463 and cg06979108 (*NOX4*) markers revealing a MAD from chronological age of 4.7 years (Lee *et al.*, 2015).

In the following year, it was proposed a specific APM for different teeth layers (cementum, dentin and dental pulp) by Giuliani *et al.* (2016). They investigated methylation data at *ELOVL2*, *FHL2* and *PENK* loci by Maldi-Tof mass spectrometry in DNA collected from 21 tooth samples (aged 17-77 years old) and obtained the highest accuracy for pulp with a median absolute difference between the estimated and chronological ages of 2.25 years, following by cementum (2.45 years) and dentin (7.07 years). These differences in the accuracy obtained in the teeth layers can be relevant for forensics, allowing the discovery of the best layers for development of more accurate APMs.

In the same year, Eipel *et al.* (2016) tested in 55 buccal swabs from healthy individuals (aged 1-85 years old) the predictive equation previously developed by

Weidner *et al.* (2014) built in 82 blood samples with methylation information of cg17861230 (*PDE4C*), cg02228185 (*ASPA*) and cg25809905 (*ITGA2B*), referred as the “3-CpG-blood-model”. Eipel *et al.* (2016) obtained a high correlation between predicted and chronological ages ($R^2 = 0.91$) however, a higher value of MAD was obtained in buccal swabs (14.6 years), comparing to the value obtained by Weidner in blood (MAD = 5.4 years). Therefore, Eipel *et al.* (2016) using the methylation information of the same CpG sites developed a specific predictive equation in the training set of 55 buccal swabs (referred as “3-CpG-swab-model”), achieving a MAD of 4.3 years and explaining 92% of the variation in age. Moreover, the evaluation of age correlation values of the three CpGs, revealed for CpG located at *PDE4C* a higher age correlation in buccal swabs and saliva compared to blood samples. The remaining sites located at *ASPA* and *ITGA2B* genes revealed higher age correlation values in blood. A simple linear regression model developed with the *PDE4C* marker explained 91% of the variation in age and allowed to obtain a MAD of 5.2 years. This value is similar to the obtained in the “3-CpG-swab-model”, suggesting that the use of only one CpG at *PDE4C* gene can be reliable for age prediction in buccal swabs. These results confirmed the tissue specificity of DNAm age-related markers, revealing that some CpGs are more suitable on certain sample types comparing with others.

Still in 2016, Freire-Aradas *et al.* (2016) investigating a total of 22 candidate genomic regions addressing 177 CpG sites, developed an APM in 725 blood samples (aged 18-104 years old) with an accuracy of 3.07 years using the seven highest age-correlated markers: *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *C1orf132* and Chr16:85395429. An online age prediction tool considering the methylation data of the seven CpGs was freely available in: <http://mathgene.usc.es/cgi-bin/snps/processmethylation.cgi>.

Zubakov *et al.* (2016) evaluating biomolecular and epigenetic approaches observed that DNAm levels allowed a higher accuracy in comparison with the other methods, as telomere length, mRNA and sjTREC. In their study, the best epigenetic model included eight DNAm markers, with five CpGs from the *ELOVL2* gene. The model developed with methylation information captured in 216 blood samples from individuals aged 4-82 years old (103 females; 113 males) allowed to obtain a MAD of 5.089 years. Another contribution in that year was proposed by Hamano *et al.* (2016) investigating methylation changes of *ELOVL2* and *FHL2* loci in 22 blood samples from living individuals and 52 blood samples from deceased individuals (aged 0-95 years old) using the methylation-sensitive high resolution melting (MS-HRM) method. In the training set of 74 samples, the developed model explained 83% of the variation in age (corrected $R^2 = 0.83$) with a MAD of 7.44 years. Moreover, this study also reports similar distributions of DNAm patterns between blood samples from living and deceased individuals.

In 2017, Cho *et al.* (2017) and Thong *et al.* (2017) replicated the study of Zbieć-Piekarska *et al.* (2015b) in blood samples from two different populations. Cho *et al.* (2017) evaluating DNAm levels of 32 CpGs located at *ELOVL2*, *FHL2*, *C1orf132*, *TRIM59* and *KLF14* genes observed age-correlated values higher than 0.455 in all the CpGs in 100 blood samples from Korean individuals (aged 20-74 years old) by pyrosequencing. The strongest age correlation value was observed for *ELOVL2* (Chr6:11044628) (0.921) which was not investigated by Zbieć-Piekarska *et al.* (2015b). For the remaining genes, in both Polish and Korean populations, the strongest age correlation sites per locus were the same: *C1orf132* (Chr1:207823681), *TRIM59* (Chr3:160450189), *KLF14* (Chr7:130734355), and *FHL2* (Chr2:105399282). In their study, Cho and collaborators (2017) applied to a Korean population sample the available predictive equation proposed by Zbieć-Piekarska *et al.* (2015b), with *ELOVL2*

(Chr6:11044634), *C1orf132* (Chr1:207823681), *TRIM59* (Chr3:160450199), *KLF14* (Chr7:130734355) and *FHL2* (Chr2:105399288), obtaining a very strong correlation value between chronological and predicted ages (0.940), with a MAD of 4.18 years. This value is similar to the reported by Zbieć-Piekarska *et al.* (2015b) (MAD = 3.4 years). Additionally, Cho *et al.* (2017), with the best-selected sites in Korean dataset for *ELOVL2* (Chr6:11044628), *C1orf132* (Chr1:207823681), *TRIM59* (Chr3:160450189), *KLF14* (Chr7:130734355) and *FHL2* (Chr2:105399282) genes, built an APM with a MAD of 3.34 years. This MAD value is lower comparing with the MAD obtained when applying the online equation of Zbieć-Piekarska *et al.* (2015b), showing that the selection of specific sites for each population can lead to an improvement in the model accuracy. Moreover, Cho *et al.* (2017) tested the association of DNAm with another type of age biomarker (sjTREC) showing the potential of multidisciplinary approaches to decrease the age prediction error in the older individuals.

Thong *et al.* (2017) evaluated DNAm levels of several CpGs located at *ELOVL2*, *FHL2*, *C1orf132*, *TRIM59* and *KLF14* genes in 145 blood samples from Singapore population (aged 3-80 years old) by pyrosequencing methodology. Strong age correlation values were observed in all 32 CpG sites ($R > 0.706$), and the highest age correlation value was obtained in the position C5 (Chr6:11044642) located at *ELOVL2* ($R = 0.957$; corrected $R^2 = 0.915$). The developed APM with *ELOVL2* (Chr6:11044642), *TRIM59* (Chr3:160450189) and *KLF14* (Chr7:130734357) revealed an accuracy of 3.3 years, explaining 95.5% of the variation in age. Thong *et al.* (2017) applied their model to bloodstains extracted immediately (day 0) and extracted after 21 days (day 21) observing no significant differences between the obtained MAD values, demonstrating the stability of DNAm levels in bloodstains, in concordance with previous reports made by Zbieć-Piekarska *et al.* (2015a) and Huang *et al.* (2015). In addition, Thong and collaborators

(2017) validated their APM with *ELOVL2*, *TRIM59* and *KLF14* in an independent set of blood samples and tested/applied the available predictive equation of Zbieć-Piekarska *et al.* (2015b) in the same independent validation set obtaining MAD values of 5.0 years and 4.8 years, respectively. Both MAD values were similar, however the two APMs selected different CpGs from the same locus and included different number of CpGs: five CpGs in Zbieć-Piekarska *et al.* (2015b) vs. three CpGs in Thong *et al.* (2017). Considering Cho *et al.* (2017), Zbieć-Piekarska *et al.* (2015b) and Thong *et al.* (2017) studies, the best CpG in each population set was always located at *ELOVL2* gene. In addition, for the *C1orf132*, *KLF14* and *FHL2* genes it was always the same CpG position showing the highest age correlation value. However, few little differences were observed in age correlation values for each CpG, possible reflecting population-specific differences.

Spólnicka *et al.* (2017) tested DNAm of *ELOVL2*, *C1orf132*, *FHL2*, *TRIM59* and *KLF14* genes using the pyrosequencing method in 190 blood samples (aged 12-76 years old) from three groups of individuals with medical conditions (Graves's disease and early or late onset Alzheimer's disease), and compared DNAm information with healthy controls (425 samples, aged 2-75 years old). *ELOVL2* and *C1orf132* markers revealed unchanged age prediction accuracy in all the groups of individuals with medical conditions and consequently they seem promising age-correlated genes to be applied for age predictions in forensic fields.

Also, during 2017, Hong *et al.* (2017) evaluated the methylation values of seven CpG sites - cg18384097, cg00481951, cg19671120, cg14361627, cg08928145, cg12757011, and cg07547549 located at *PTPN7*, *SST*, *CNGA3*, *KLF14*, *TSSK6*, *TBR1*, and *SLC12A5* genes, respectively, in 226 saliva samples (aged 18-65 years old) by SNaPshot methodology. A high age correlation value between chronological and

predicted ages (0.945) was obtained in the training set of 113 samples, with a MAD from chronological age of 3.13 years. Another APM developed by Hong *et al.* (2017) with six CpGs located at *SST*, *CNGA3*, *KLF14*, *TSSK6*, *TBR1* and *SLC12A5* (without cg18384097 located at *PTPN7* locus, which is a cell type-specific CpG), revealed a MAD of 4.1 years. Comparing both models (APM with seven CpGs and six CpGs), it is observed that the addition of cg18384097 allowed an improvement of the age prediction.

More recently, Lee *et al.* (2018) applied their previously developed APM in semen samples (Lee *et al.*, 2015) to 12 independent semen samples (aged 24-57 years old) and 19 different forensic samples containing semen (aged 17-48 years old), revealing the usefulness of their model in forensic samples. Using the 12 samples, they obtained a high age correlation value between chronological and predicted ages (0.970) with MAD of 4.8 years. Applying their APM to 19 forensic samples, a strong correlation value between chronological and predicted ages was achieved (0.773), with a MAD from chronological age of 5.2 years (Lee *et al.*, 2018).

Also in 2018, Horvath *et al.* (2018) developed the “skin & blood clock” model focused on 391 CpGs, which were analyzed in several tissues, including buccal swabs, saliva, blood, endothelial cells and skin. The “skin & blood clock” model demonstrated highly correlation value between predicted and chronological ages in blood ($R = 0.98$) with a median absolute deviation between predicted and chronological ages of 2.5 years. Moreover, when applied to trabecular bone samples, previously collected and used in Horvath *et al.* (2015), this model revealed a correlation value between predicted and chronological ages of 0.82. Of note, this model revealed the influence of lifestyle choices and other environmental conditions for age predictions based on DNAm (Horvath *et al.*, 2018).

In the same year, Naue *et al.* (2018) through the massive parallel sequencing (MPS) methodology, evaluated DNAm in 144 different biological samples from 29 deceased individuals using 13 loci previously tested in 324 blood samples from living individuals (Naue *et al.*, 2017), including *ELOVL2*, *TRIM59* and *KLF14* markers. Strong age correlation values were obtained for: *ELOVL2* cg16867657 (Chr6:11044644) in blood, brain, buccal swabs and muscle ($R \geq 0.76$); *TRIM59* cg07553761 (Chr3:160450189) in blood, brain and buccal swabs ($R \geq 0.81$); and for the position located at *KLF14* (Chr7:130734357) in brain and buccal swabs ($R \geq 0.70$). For bone samples, these markers revealed moderate age correlation values ($0.51 \leq R \leq 0.61$). This study showed how DNAm specific-tissue variability is important for the selection of age markers suitable for each tissue type and consequently for the development of accurate tissue-specific APMs.

Freire-Aradas *et al.* (2018) tested DNAm levels by EpiTyper analysis in blood samples from 180 young Europeans (aged 2-18 years old) developing a model for children with CpGs located at *KCNAB3*, *EDARADD*, *ELOVL2*, *CCDC102B*, *C1orf132* genes and position CR_23_CpG3 (no gene), with a median absolute error of 0.94 years and 77.8% of correct predictions.

To the best of our knowledge, Shi *et al.* (2018) made a first study combining anthropological and epigenetic approaches. They evaluated 124 Chinese children (aged 6-15 years old) through anthropological methods (skeletal and dental ages) of which 48 children were also evaluated by epigenetic methods (based on DNAm markers). Evaluating a total of 485.577 CpGs using the Illumina Human Methylation 450 Bead Chip, they validated by droplet digital PCR five age-correlated markers (*DDO*, *PRPH2*, *DHX8*, *ITGA2B* and one unknown gene with the Illumina ID number of 22398226) that were used to build APMs with high accuracy of age prediction in boys and girls, with

values of MAE (mean absolute error) = 0.47 years and MAE = 0.33 years, respectively. This study also put on top the use of the multidisciplinary approach for age assessment.

Jung *et al.* (2019) evaluated DNAm of only five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*, and *TRIM59* genes using a SNaPshot assay in 448 samples (aged 18-74 years old), including blood, saliva and buccal swabs. The promising multi-tissue APM developed in 300 samples showed a very strong correlation value between predicted and chronological ages ($R = 0.950$) and a MAD from chronological age of 3.553 years. The authors defended that some markers (as *TRIM59* and *ELOVL2*) could functioning as multi-tissue markers, showing a strong age correlation either in blood, saliva and buccal swabs; other genes were more tissue-specific (as *FHL2* and *C1orf132*), emphasizing the need to test markers in each tissue type.

Also in this year, Gopalan *et al.* (2019) developed, for the first time, a specific APM for bone samples: a “37 bone clock CpGs” based on methylation information of CpGs located, among other genes, at *TRIM59*, *ELOVL2* and *KLF14*. They used methylation data from bone samples from living individuals previously published from Illumina Human Methylation 450K array and data collected for their study from deceased individuals from Illumina Human Methylation EPIC array methodology. This study reveals a RMSE of 4.9 years in all the samples analyzed and supports the idea that *ELOVL2* is a powerful gene to be applied for age prediction in bones.

Most recently, Lee *et al.* (2020) used 66 femoral bone samples (aged 31-96 years old) for DNAm analysis. They evaluated through the SNaPshot of Jung *et al.* (2019), 34 samples (23 males and 7 females; and 4 samples for a preliminary test) and the remaining 32 bone samples (28 males and 4 females) through Infinium HumanMethylation EPIC BeadChip array, using the “skin & blood clock” previously developed by Horvath *et al.* (2018). For the 30 bone samples using the SNaPshot assay of Jung *et al.* (2019), the

obtained age correlation values for *TRIM59* (cg07553761; Chr3:160450189) and *ELOVL2* (Chr6:11044628) were $R = 0.434$ and $R = 0.415$, respectively. From the 32 bone samples, only 12 with high-quality data were analyzed using the “skin & blood clock”.

Page | 42

Between the 391 CpGs examined, 38 CpGs (including cg06639320 and cg22454769 from *FHL2*; cg09809672 from *EDARADD*; cg21572722 and cg16867657 from *ELOVL2*; cg20426994 from *KLF14* and cg07553761 from *TRIM59* genes) revealed a strong age correlation value in bone samples ($R > 0.70$). Predicting age using the 12 bones (11 males; 1 female) allowed to obtain a correlation value between predicted and chronological ages $R = 0.964$, with a MAD from chronological age of 6.4 years.

A recent study proposed by Márquez-Ruiz *et al.* (2020) tested DNAm levels of specific CpGs located at *ELOVL2*, *ASPA* and *PDE4C* genes by bisulfite pyrosequencing using 65 tooth samples from individuals aged 15-85 years old. They observed a positive and moderate age-correlated value for *ELOVL2* CpG9 (Chr6:11044642) ($R = 0.595$) and a positive age correlation value near to moderate for *PDE4C* CpG1 (Chr19:18343916) ($R = 0.465$). The developed APM with CpGs from these two loci presented a MAE of 5.08 years.

Finally, Koop *et al.* (2020) evaluated by pyrosequencing DNAm changes of one CpG from *PDE4C* gene in 215 buccal swabs from living and deceased individuals. The developed one-CpG APM in 71 living individuals revealed higher age correlation value ($r^2 = 0.87$). The validation of this APM in an independent set of 71 swab samples of living individuals revealed moderate accuracy with a MAD of 7.8 years. The application of this predictive equation to 52 swab samples of deceased individuals allowed to obtain a moderate accuracy with a MAD = 9.1 years. In addition, no differences in age estimation was observed for deceased samples with different stages of body decomposition.

Additionally, in recent years, some studies have defended the construction of DNAm biological clocks, distinct from the previous chronological clocks (Field *et al.*, 2018; Horvath and Raj 2018). To date, as the best of our knowledge, only six studies have tried the development of biological clocks using DNAm features linked to many factors as risk of mortality, diseases and other age-clinical conditions (Yang *et al.*, 2016; Zhang *et al.*, 2017b; Levine *et al.*, 2018; Youn and Wang, 2018; Lu *et al.*, 2019a; Lu *et al.*, 2019b). These biological clocks focus on inter-individual divergence, being more related with the epigenetic drift (Field *et al.*, 2018; Horvath and Raj 2018). Consequently, these clocks can be more advantageous in clinical research than in forensic investigations, in which the estimation of individual's chronological age is the paramount issue (Field *et al.*, 2018).

In summary, in the last decade, many investigations in forensic age prediction have been reported using different methodological approaches to assess DNAm levels that are considered the most promising age prediction biomarker. Several APMs were developed in several specific tissue types, essentially in blood samples from living individuals, and others were suggested across multiple tissues. Similar age predictions accuracies were consistently obtained based on these different methodological approaches using several markers, different tissue types or different statistical methods. The *ELOVL2* gene has been widely investigated, showing to be the most powerful age prediction marker to date. Using many different sources of DNA, such as saliva, blood, teeth, buccal cells, liver, brain, bone and old bloodstains, *ELOVL2* CpGs demonstrate the higher rates of age association. Meanwhile, the development of multi-tissue APMs has arisen. However, it is necessary to clarify what are the best multi-tissue or tissue-specific age markers. It is important to discovery and select the specific markers that can show similar accuracy in several types of samples using the same methodology for development

of APMs applied to forensic casework. After that, DNAm levels can become an accurate and reliable method to be applied for many contexts in forensic age estimations.

4. Challenges for application of DNAm age estimation in forensic contexts

For the application of DNAm age at forensic contexts, several aspects should be considered:

i) the fact that common forensic samples contain a small amount of DNA, low quality of DNA or degraded DNA; this can be a big challenge for forensic casework when we deal with DNAm analysis requiring, for instance, bisulfite conversion (in which occurs DNA loss);

ii) the APMs that have already been developed for each specific tissue type present in the forensic case and the selection of the most accurate APM;

iii) the simplicity, cost and practicality of each method, including the necessity of specific and available equipment, and time consuming in solving forensic questions (Jung *et al.*, 2017);

iv) the restriction of the number of CpGs to 10 or less, in order to be practicable for forensic casework (Alsaleh *et al.*, 2017); however, it is known that the development of multi-tissue APMs requires the inclusion of a larger number of CpG sites.

Additionally, DNAm levels can be affected by intrinsic influences (as gender, aging or ancestry) or by several environmental factors (including lifestyle, disease, alcohol consumption or social environments) (Kader and Ghai, 2015). A recent study by Fiorito *et al.* (2019) revealed the impact of several socioeconomic factors (as education)

and lifestyle choices (as tobacco and alcohol consumption) in DNAm clocks, as the Hannum clock and the Horvath clock (Fiorito *et al.*, 2019).

Some of the most important factors that should be considered when using DNAm age for forensic purposes are addressed below.

4.1. Differences between predicted and chronological ages with aging

According to literature, younger individuals show lower values of MAD between predicted and chronological ages, reflecting a high accuracy in the APMs, comparing to older ages (Horvath, 2013; Zbieć-Piekarska *et al.*, 2015b; Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Thong *et al.*, 2017; Naue *et al.*, 2017; Pfeifer *et al.*, 2020).

The observed lower accuracy of the APMs with aging can reflect the higher differences between biological and chronological ages of individuals with the increase of age. This can be related to the accumulation of specific alterations in DNAm patterns of each individual with aging due the stochastic or environmental factors, being accepted as the epigenetic drift contribution (Jones *et al.*, 2015; Tan *et al.*, 2016; Freire-Aradas *et al.*, 2017).

4.2. Specificity of each age-correlated marker to diseases

DNAm patterns of some genes can be affected by some diseases, and consequently the accuracy of the epigenetic clocks can be influenced. Spólnicka *et al.* (2017) suggested that some genes can keep their power as age predictors regardless of the presence of diseases, but other markers seem to be influenced by some diseases. This supports the idea that each marker can reveal a specificity and/or sensibility to each disease, being more or less accurate for measurement of chronological age (Bacalini *et*

al., 2017; Spólnicka *et al.*, 2017). Thus, for development of APMs highly accurate in forensic casework the influence of certain common diseases in DNAm of specific markers should be considered. In addition, Bell *et al.* (2019) proposed the construction of disease-specific APMs based on DNAm evaluation as future direction in forensic field.

4.3. Tissue specificity

In recent years, many tissue-specific APMs have been developed for blood, saliva, buccal swabs, bone and teeth. Moreover, few researchers built multi-tissue APMs that can be applied to several combined types of samples as made by Horvath (2013) with high prediction accuracy. However, Horvath (2013) included 353 CpGs to capture age-related changes across different tissue types. The inclusion of a larger number of markers improves the prediction accuracy but can be a disadvantage in forensic casework. Moreover, it has been observed that there are tissues, in which age estimation leads to a high error of estimate. This can be explained by tissue specificity of each age-correlated marker. Also, Eipel *et al.* (2016) demonstrated that the specific predictive equation developed by Weidner *et al.* (2014) for blood samples cannot be applied to buccal swabs with the same accuracy in age predictions. In any case, it seems that there are some markers with the ability to predict age across several tissues (and could be assigned as multi-tissue markers), as the *ELOVL2* locus (Hannum *et al.*, 2013; Jung *et al.*, 2019), while others are more tissue-specific, such as *FHL2* and *C1orf132* genes (Jung *et al.*, 2019). Consequently, studies to test individual age-correlated markers in specific tissues and to adapt specific APMs (developed for a specific tissue type) to other tissue types are needed.

4.4. Specificity of population groups

Population-specific patterns of DNAm can occur in several CpGs over the genome. Fraser *et al.* (2012) using a 27.000 CpG site microarray platform observed differences between African and European populations. Heyn *et al.* (2013) investigated around 450.000 CpGs in three populations (Caucasian-American, African-American, and Han Chinese-American) observing DNAm differences in several CpGs. They suggested that these differences allowed the separation of populations and could explain some patterns of the human variation.

For highly accurate age estimation, it is necessary to test the predictive equations that have already been developed for specific-population groups in other populations to address potential differences in DNAm levels at the same CpG positions. This has been done by few groups as Cho *et al.* (2017) in Koreans and Thong *et al.* (2017) in Singaporean individuals applying the APM of Zbieć-Piekarska *et al.* (2015b) developed in Polish (MAD = 3.4 years). The model accuracy remains similar when applying the predictive equation developed in Polish individuals (Zbieć-Piekarska *et al.*, 2015b) to Korean individuals (MAD = 4.18 years) (Cho *et al.*, 2017) and to Singaporean individuals (MAD = 4.8 years) (Thong *et al.*, 2017). Nevertheless, Cho *et al.* (2017) observed an improvement of the accuracy when developed specific APMs for Koreans.

Additionally, a study by Fleckhaus *et al.* (2017) applied the two previous predictive equations proposed in German people by Eipel *et al.* (2016) and in Polish by Zbieć-Piekarska *et al.* (2015b) to three independent population groups from Middle East, West Africa and Central Europe. They reported the same strength of change in DNAm levels across the different CpGs for the three populations, but observed a slight high accuracy in Middle East population comparing with the two other populations. These

results suggest that APMs may have to be adjusted for different populations, which leads to the increase of the accuracy.

4.5. Sex differences

Another factor that could influence age estimation based on DNAm levels is the putative effect of the sex, but no consistent results were obtained to date. Boks *et al.* (2009) evaluating 1.505 CpGs from more than 800 genes by Illumina GoldenGate Methylation assay demonstrated that the sex could be a factor for DNAm changes at specific loci across the genome. When testing DNAm changes of specific age-related CpGs considering sex as a possible influencer, some authors observed few differences between males and females (Weidner *et al.*, 2014; Zbieć-Piekarska *et al.*, 2015b; Daunay *et al.*, 2019), despite not being relevant in the accuracy of age predictions. DNAm levels of *ELOVL2* and *PDE4C* genes seem to be affected by sex in few studies (Weidner *et al.*, 2014; Zbieć-Piekarska *et al.*, 2015b), but no statistically significant sex differences were observed in other studies for the two loci (Bekaert *et al.*, 2015a; Freire-Aradas *et al.*, 2016, 2018; Márquez-Ruiz *et al.*, 2020). Accordingly, some studies observed no significant sex influence in DNAm levels of other age-correlated markers (Koch and Wagner, 2011; Huang *et al.*, 2015; Bekaert *et al.*, 2015a; Freire-Aradas *et al.*, 2016, 2018; Márquez-Ruiz *et al.*, 2020). In any case, considering these reports, the possible influence of sex in DNAm of age-related markers should be explored whenever possible in each developed APM.

Chapter 2. Objectives

Chapter 2. Objectives

Although many CpG markers have been identified showing high correlations with chronological age in previous studies, validation of methodologies and markers should be made to develop accurate APMs, potential useful in forensic analysis. Because of the highly tissue-specific nature of DNAm, the development of APMs for particular types of tissues should also be tested. In addition, it is possible that few markers could reveal high accuracies across different tissue types allowing the development of multi-tissue APMs.

Most forensic APMs have been developed based on blood DNA samples of healthy individuals, but other tissues are now also being explored. For example, four studies focused on the possible DNAm changes in blood samples from deceased individuals, which can be a relevant source of DNA for forensic age estimation models (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Naue *et al.*, 2018; Pfeifer *et al.*, 2020). In the same way, a number of studies used tooth samples (Bekaert *et al.*, 2015a; Giuliani *et al.*, 2016; Márquez-Ruiz *et al.*, 2020) or bone samples (Horvath *et al.*, 2015, 2018; Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020) for DNAm age prediction. Bones and teeth are valuable sources of DNA being, very often, the last evidence of an individual in cases of forensic identification.

Considering the aforementioned review addressing the highly age-correlated genes for the development of APMs based on DNAm levels in several tissue types, a number of genes were selected for the present study: the seven previously known and

validated age-associated genes *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C*, *C1orf132*/*MIR29B2CHG*, *KLF14* and *TRIM59*.

In the top of this list, being one of the most investigated age predictors markers, is the *ELOVL2* (ELOVL Fatty Acid Elongase 2) gene, located on the chromosome 6p24.2. CpG sites from the *ELOVL2* gene has been continually included in several APMs using different tissue types as blood, buccal swabs, saliva, bones and tooth samples (Garagnani *et al.*, 2012; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Giuliani *et al.*, 2016; Park *et al.*, 2016; Freire-Aradas *et al.*, 2016; Hamano *et al.*, 2016, 2017; Cho *et al.*, 2017; Thong *et al.*, 2017; Spólnicka *et al.*, 2017; Bacalini *et al.*, 2017; Naue *et al.*, 2017, 2018; Jung *et al.*, 2019; Gopalan *et al.*, 2019; Daunay *et al.*, 2019; Márquez-Ruiz *et al.*, 2020; Lee *et al.*, 2020; Pfeifer *et al.*, 2020).

The remaining genes, including *FHL2* (Four And A Half LIM Domains 2), located on chromosome 2q12.2, *EDARADD* (EDAR Associated Death Domain), located on chromosome 1q42.3, *PDE4C* (phosphodiesterase 4C, cAMP specific), located on chromosome 19p13.11, *C1orf132* (chromosome 1 open reading frame 132), located on chromosome 1q32.2, *TRIM59* (Tripartite motif containing 59), located on chromosome 3q25.33, and *KLF14* (Kruppel-like factor 14), located on chromosome 7q32.3, are other promising age-correlated genes that have been also investigated in several studies using different biological samples including blood, buccal swabs and saliva (Garagnani *et al.*, 2012; Florath *et al.*, 2014; Zbieć-Piekarska *et al.*, 2015b; Hamano *et al.*, 2016; Giuliani *et al.*, 2016; Freire-Aradas *et al.*, 2016; Cho *et al.*, 2017; Thong *et al.*, 2017; Spólnicka *et al.*, 2017; Bacalini *et al.*, 2017; Jung *et al.*, 2019; Daunay *et al.*, 2019; Pfeifer *et al.*, 2020; Koop *et al.*, 2020). In particular, *FHL2*, *EDARADD* and *PDE4C* genes were included in several studies for DNAm age prediction using tooth samples (Bekaert *et al.*, 2015a; Giuliani *et al.*, 2016; Márquez-Ruiz *et al.*, 2020) and *TRIM59*, *KLF14*, *C1orf132* and

FHL2 genes revealed its valuable utility for DNAm age estimation in bones (Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020).

For measurement of DNAm levels among selected CpG sites from the chosen candidate genes, we used two different methodologies: the bisulfite polymerase chain reaction (PCR) Sanger sequencing and the SNaPshot multiplex system. Both methods used herein provide a semi-quantitative measure of DNAm levels at CpG sites, but have shown to be efficient and economical alternative tools to other widely used methodologies. For example, the bisulfite Sanger sequencing was shown to have similar linearity and accuracy than pyrosequencing analysis (Jiang *et al.*, 2010; Parrish *et al.*, 2012), which is one of the most currently used methods for DNAm age estimation. The multiplex methylation SNaPshot is a simple, efficient and cost-effective way to determine simultaneously DNAm levels of different individual target CpG sites and has been currently used in forensic DNAm analysis.

That said, the main aim of this study was to develop several APMs based on the previously known and validated age-associated genes *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C*, *C1orf132/MIR29B2CHG*, *KLF14* and *TRIM59* by evaluating the correlation between DNAm levels and chronological age in several types of biological samples from Portuguese individuals and using two different methodologies: bisulfite sequencing, addressing methylation levels in genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132*, and the SNaPshot assay of Jung *et al.* (2019) to address five CpGs located in genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59*.

From this main goal, several specific objectives were considered:

1. to develop simple and multiple APMs for blood samples from both living and deceased individuals (collected during autopsies and collected from identified Bodies Donated to Science, BDS), teeth from living and deceased individuals (collected from BDS), bones (collected during autopsies and from BDS) and buccal swabs from living individuals.
2. to examine possible differences in methylation patterns between blood samples from living and deceased individuals;
3. to investigate methylation differences according to aging, sex, ancestry groups and PMI (*postmortem* interval);
4. to evaluate the accuracy of bisulfite Sanger sequencing and SNaPshot methodologies by comparing methylation data obtained with the two methodologies;
5. to develop multi-tissue APMs using different combination sets of biological samples;
6. to evaluate DNAm patterns and to estimate age in dry bones from The 21st Century Identified Skeletal Collection (ISC/XXI) or *Coleção de Esqueletos Identificados do Século XXI* (CEI/XXI).

Chapter 3. Sample and design research

Chapter 3. Sample and design research

1. Types of samples, collection, handling and storage

Several types of biological material from living and deceased individuals were collected (blood, bone, tooth and buccal swabs). Sample collection procedures and stored conditions were according to each tissue type. Store conditions were chosen to avoid DNA degradation (normally using low temperatures and humidity). Improper handling of a sample can lead to contamination, sample degradation or even misinterpretation of the results.

1.1. Blood samples

A total of 144 blood samples was collected from living and deceased individuals:

i) peripheral blood samples from 71 healthy living individuals of Portuguese ancestry (45 females, 26 males; aged 1-95 years old) were collected from users of *Biobanco - Hospital Pediátrico de Coimbra* and other hospitals. Data from each individual (age, sex, clinical conditions) were collected in the written consent form;

ii) blood samples from 73 deceased individuals (15 females, 58 males; aged 24-91 years old) were collected during routine autopsy (12 females, 52 males), after consulting RENNDA (*Registo Nacional de Não Dadores*) in *Serviço de Patologia Forense da Delegação do Centro do Instituto Nacional de Medicina Legal e Ciências Forenses* (INMLCF), and from Bodies Donated to Science (BDS) (3 females, 6 males), before the embalming method with Thiel (Eisma *et al.*, 2013) in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* (Annex I, BDS form

example). BDS have an indispensable role in advancing of Medical Science in general. They are essential for anatomical studies in medical schools, as well as in scientific and technical research that leads to the progress of medical and surgical treatments, according to the Portuguese legislation that regulates the dissection of cadavers (*Decreto-Lei n.º 274/99, de 22 de Julho*). A table with the biological features and the type of biological sample collected from each BDS was present in Annex II. Data from each individual (age-at-death, sex, cause of death, diseases indicators) were assessed in medical reports provided by the legist doctor (autopsies) or by the *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* (BDS).

All blood samples from dead bodies were collected within five days after death. The available *antemortem* data indicate no evidence of cancer, which could affect the methylation status (Hannum *et al.*, 2013).

All blood samples from living and deceased individuals were collected using a syringe, transferred to a clean sterile storage tube that contains anticoagulant (EDTA) and were frozen at -20°C in vertical position, until DNA extraction (**Figure 3.1**). The processing of blood samples occurred in *Laboratório de Genética Humana* (LGH) from CIAS (*Centro de Investigação em Antropologia e Saúde*).



Figure 3.1: Blood samples from deceased individuals collected during autopsies in INMLCF.

1.2. Bone samples

A total of 62 bone samples was collected for this study:

i) a set of 31 bone samples from deceased individuals collected after consulting RENNDA, during autopsy in *Serviço de Patologia Forense das Delegações do Centro e Sul do INMLCF* (5 females, 26 males; aged 26-81 years old). These bones were collected within five days after death and the available data from each individual (age, sex, cause of death, diseases indicators) were assessed in medical reports provided by the legist doctor. We excluded individuals with known diseases or other clinical conditions that could influence DNAm levels;

ii) a set of 22 bones from BDS (12 females, 10 males; aged 49-93 years old) (Annex I, BDS form example) collected after the embalming method with Thiel (Eisma *et al.*, 2013), in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* and in *Faculdade de Medicina da Universidade de Coimbra*. Bones from BDS were collected between 1 to 4 years after death (Annex II). For each individual the available information including, sex, age-at-death, state of the body) was provided by the *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*;

iii) nine dry bones (7 females, 2 males; aged 38-92 years old) from The 21st Century Identified Skeletal Collection (ISC/XXI) or *Coleção de Esqueletos Identificados do Século XXI* (CEI/XXI) were collected after approval by the *Laboratório de Antropologia Forense* (LAF) from CEF (*Centro de Ecologia Funcional*). The CEI/XXI arises from a collaboration in 2007 between the Department of Anthropology of the University of Coimbra and the City Council of Santarém (responsible for the cemetery from which the skeletons originate) to carry the unclaimed skeletons to the University for investigation purposes. The City Council of Santarém made available copies of

certificates of death of each individual, and other relevant information, as name, age-at-death, sex, nationality, between other data (Ferreira *et al.*, 2014).

Page | 60

The bone samples from deceased individuals collected during autopsies were involved in aluminum foil, with label identification (case number, age and sex), and individualized in plastic bags (**Figure 3.2A**). The recommendation should be the individualization in paper bags, however as the bones were frozen at -80°C , we choose plastic bags. Bones samples from BDS were collected in small plastic recipients with identification number and then freeze at -80°C (**Figure 3.2B, C**). Dry bones samples from CEI/XXI were individualized in plastics bags and were kept at room temperature (**Figure 3.3**).

Bone collection took place from December 2018 to January 2020. The processing of bone samples occurred in LGH from CIAS and in *Serviço de Genética e Biologia Forenses* (SGBF) from INMLCF, according to the recommended guidelines.

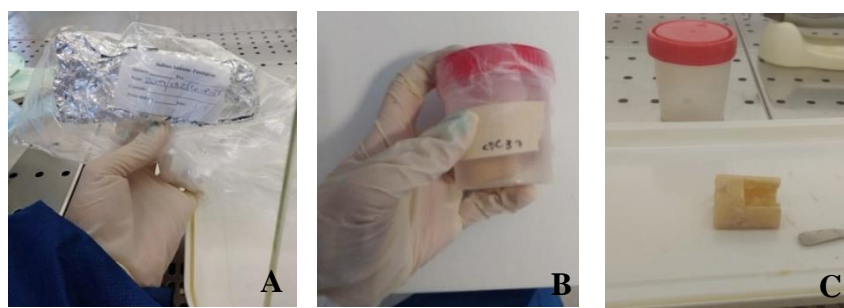


Figure 3.2: Bones. A) Bone sample collected from one deceased individual during autopsy in INMLCF; B) Bone sample collected from BDS; C) Bone sample collected from BDS, the photo was captured after bone cut.



Figure 3.3: Dry bone sample from CEI/XXI, the photo was captured after bone cut.

1.3. Tooth samples

A total set of 31 tooth samples were used in this research. Twenty-three tooth samples from living individuals (16 females, 7 males; 27-94 years old) collected in dentist offices, after written informed consent, and 8 tooth samples from deceased individuals collected from BDS (5 females, 3 males; 48-88 aged years old) in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* before the embalming method of the body (Annex I, BDS form example; Annex II). The collection of teeth before the embalming ensures the control of any possible influence related to the process of conservation. Available data such as age, sex and clinical conditions were collected with the informed consent for living individuals, and were provided by the *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* for deceased individuals. Tooth samples from living and deceased (BDS) individuals were collected in clean tubes with absolute ethanol and kept at room temperature until analysis (**Figure 3.4**). The processing of tooth samples occurred in SGBF from INMLCF, according to the recommended guidelines, and in LGH from CIAS.

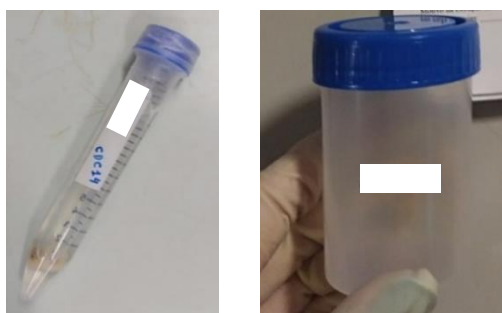


Figure 3.4: Tooth samples in plastic recipient.

1.4. Buccal swabs

A set of 39 buccal swabs from living individuals of Portuguese ancestry (16 males, 23 females; aged 3-86 years old) was collected with a sterile brush (Sarstedt, Nümbrecht,

Germany) (**Figure 3.5**). Buccal swabs were stored under refrigeration (short periods) or frozen at -20°C (for long-term storage) (Goodwin *et al.*, 2007). The processing of buccal swabs occurred in LGH from CIAS. Data of each individual such as age, sex and clinical conditions were collected with the informed consent form (Annex I, buccal swab consent).



Figure 3.5: Buccal swabs from living individuals.

2. Ethical or institutional guidelines

The study protocol of this doctoral research was approved by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra*, process nº 038-CE-2017 (Annex III, Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra*). For blood, buccal swabs and tooth samples, from living individuals, written informed consent was previously obtained from adult participants and from children's parents, under the age of 18 years. In addition, for blood and bone samples collected during autopsies we request for approval by the Ethical committee of the INMLCF.

3. Laboratory methods

3.1. DNA extraction

3.1.1. Blood samples

Genomic DNA extraction from blood samples was performed in LGH from CIAS, using the *QIAamp DNA Mini Kit* (Qiagen, Hilden, Germany), according to the instructions of the manufacture. Briefly, the procedure consisted in four steps: i) lysis, consisting in addition of 20 μ l of proteinase K (Qiagen) (10 mg/ml) and 200 μ l of Lysis Buffer AL to 200 μ l of blood in a 1.5 ml microcentrifuge tube; ii) binding DNA in the lysate to a silica-based membrane in a QIAamp Mini spin column; iii) washing contaminants with two buffers (AW1 and AW2); and iv) elution of DNA (50 μ l) from the membrane in a clean 1.5 ml microcentrifuge tube with buffer AE. The DNA solution was stored at 4°C or -10°C (see Annex IV, for the complete protocol).

3.1.2. Bone and tooth samples

DNA extraction from bones and tooth samples was made in INMLCF, according to standard guidelines. All sample manipulations were performed in laminar flow chambers equipped with filters and UV lights (Workstation I and Workstation II). In workstation I, we made a pre-treatment of bone and tooth samples (Annex IV, for guide with steps and illustrations). Briefly, after decontamination of laminar flow chambers, the bone fragment or root tooth were collected using a chainsaw; and put in commercial bleach (sodium hypochlorite, 10%) during 5 minutes. Following the bone and tooth fragments were put in distilled water for another 5 min to remove residual bleach. A drill was used to remove other exogenous contaminants. After, we make some cuts (around 0.5 x 0.5 cm) in the fragment and it was put in a vial for grinding (**Figure 3.6**). In a SPEX

Sample Prep Freezer/Mill 6770 with liquid nitrogen, the bone and tooth fragments were submitted to several temperatures until a fine powder is obtain. The powder can be used immediately for DNA extraction, or kept at -80 °C.



Figure 3.6: Bone fragments in vial before grinding.

In a workstation II (**Figure 3.7A**), we used a semi-automatic protocol with *PrepFiler Express BTA™ Forensic DNA Extraction kit* (Applied Biosystems, Foster City, CA) for DNA extraction protocol (Annex IV, for guide with steps and illustrations). Briefly, we performed all the manipulations with the required reagents to prepare the extraction. Following the practical guidelines of the INMLCF and the recommendation of the manufacturer's user manual, around 50 mg fine powder (of bones or teeth) were put in a lysis tube (**Figure 3.7B**). We prepared the PrepFiler BTA™ lysis solution (each sample requires the following: 220 µl PrepFiler BTA™ lysis Buffer, 3 µl freshly prepared 1 M DTT, 7 µl Proteinase K). For each bone and tooth samples we added 230 µl of the prepared PrepFiler BTA™ lysis solution, vortex and centrifuged. We put the PrepFiler bone and tooth Lysate Tube in a thermal shaker and incubated at 56°C, 750 rpm overnight. This is the standard protocol used in INMLCF.

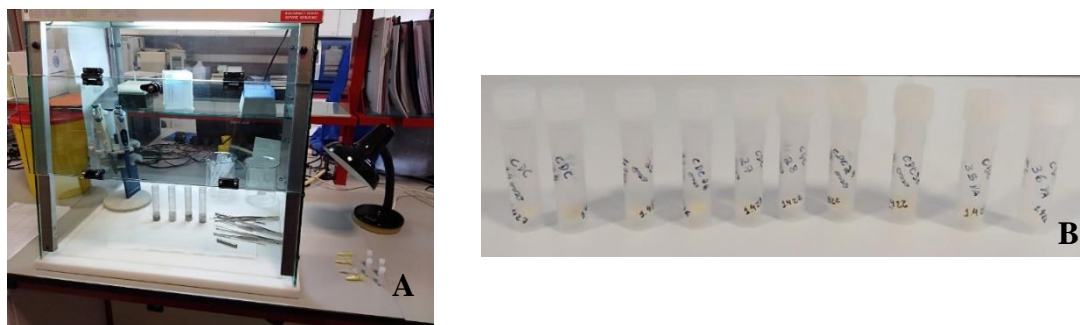


Figure 3.7: A) Workstation II; B) Bone powder before the addition of prepared PrepFiler BTA™ lysis solution.

The day after, centrifuge the PrepFiler bone and tooth Lysate Tube and the clear lysate was put in automated robot AutoMate Express™ (Applied Biosystems, Foster City, CA) Forensic DNA Extraction System (**Figure 3.8A**), which can extract 11 samples at only one run, plus positive and negative control (NTC, non-template control). The controls are always included in each run. A bloodstain collected specifically for this study according to ethical guidelines was used as a positive control.

The DNA extraction protocol is very efficient and very fast (only 30 min, after digestion overnight). Moreover, all the reagents were individualized in cartridges, only being exposed in a safe climate during DNA extraction (Zupanič Pajnič *et al.*, 2016), reducing human handling and the possibility of contamination. The isolated DNA (**Figure 3.8B**) can be stored at 4°C for up two weeks or at -20°C for longer storage.

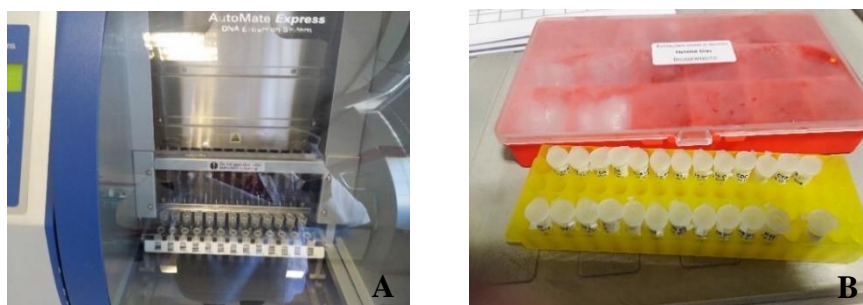


Figure 3.8: A) Robot AutoMate Express™ Forensic DNA Extraction System from INMLCF (Coimbra, Delegação do Centro); B) Isolated DNA.

For bones from CEI/XXI, an additional step was required after pre-treatment of bone fragment and before DNA extraction using *PrepFiler Express BTA™ Forensic DNA Extraction kit*. We use two different protocols with EDTA (Protocol 1 - EDTA and Protocol 2 - EDTA, described in *Chapter 4. Results and discussion: F. Dry bone samples from Coleção de Esqueletos Identificados do Século XXI-CEI/XXI*) to aid the full demineralization process, which facilitates DNA extraction. Full demineralization is considered the best method of DNA extraction from old bone material (Jakubowska *et al.*, 2012; Amory *et al.*, 2012). Demineralization shows to increase the yield of DNA obtained in old samples (Zupanič Pajnič *et al.*, 2016). Consequently, this process must be the most suitable for DNA extraction with dry bone samples from CEI/XXI, which presented higher PMI and suffered different *postmortem* influences.

3.1.3. Buccal swabs

Genomic DNA from buccal swabs was extracted using the *FavorPrep™ Genomic DNA mini kit* (Favorgen Biotech Corp, Taiwan) at LGH from CIAS, according to the instructions of the manufacture. Briefly, the procedure consisted in four steps: i) lysis - after resuspending the buccal swab in 1 ml of distilled water (**Figure 3.9**) followed by a centrifugation step, 200 µl of FABG Lysis Buffer and 20 µl of proteinase K (10 mg/ml) (VWR Life Science) were added to the pellet of buccal cells; ii) binding the lysate with DNA to a silica-based membrane in FABG columns; iii) washing contaminants with two buffers (W1 and Wash buffer); and iv) elute DNA from the membrane with 50 µl of Elution buffer. The isolated DNA was stored at 4°C or -10°C (see Annex IV, for the complete protocol).

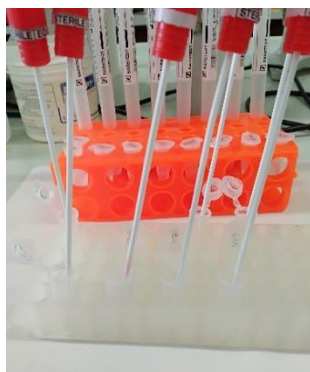


Figure 3.9: First step of DNA extraction from buccal swabs: resuspension of swabs in 1 ml of distilled water.

3.2. DNA quantification

The accurate quantification of the amount of DNA and the evaluation of its quality are a crucial step when dealing with human remains. For each tissue type, different quantification methods were chosen.

3.2.1. Blood samples and buccal swabs

DNA extraction from blood and buccal swabs were quantified in a Nanodrop spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA), according to instructions of the manufacturer. This method based on absorption of UV light, DNA absorbs light at 260 nm; other molecules (as proteins and carbohydrates) absorbs at other length (respectively 280 nm and 230 nm). Pure DNA shows the ratio of absorptions at 260 nm and 280 nm between 1.8 and 2.0. The obtained DNA concentrations for blood samples ranged between 20 to 40 ng/ μ l and for buccal swabs between 2 to 20 ng/ μ l.

3.2.2. Bone and tooth samples

DNA quantification of DNA extracts from bone and tooth samples, was made using the *QuantifilerTM Human DNA Quantification Kit* (Applied Biosystems, Foster City, CA) on the Applied Biosystems® 7500 Real-Time PCR System (HID Real-Time

PCR Analysis Software), according to the recommended guidelines (Thermo Fisher Scientific, 2018). The obtained DNA concentrations for bone and tooth samples ranged between 5 to 50 ng/ μ l.

DNA quantification assay included primers for amplify the target specific DNA and the internal PCR control (IPC); IPC template; and two TaqMan® MGB probes (one labeled with FAM™ dye, which identify the amplified sequence and the other labeled with VIC™ dye, which identify the amplified IPC) (Thermo Fisher Scientific, 2018).

DNA quantification using the *Quantifiler™ Human DNA Quantification Kit* requires the preparation of eight DNA quantification standards dilution series with known concentrations ranging from 50 ng/ μ l to 0.023 ng/ μ l. We put into each well of the reaction plate, 23 μ l of the PCR mix reaction (**Table 3.1**) and 2 μ l DNA standards or DNA samples or controls. A negative PCR control (NTC) for the quantification and the negative and positive controls of the DNA extraction were included in each run.

Table 3.1: Volume of each reagent per sample.

Reagent	Volume per reaction (μ l)
Quantifiler® Human Primer Mix	10.5 μ l
Quantifiler® PCR Reaction Mix	12.5 μ l

The IPC system (an internal control of amplification) allows the interpretation of quantification data by differentiation between true negative results or reactions affected by other factors as PCR inhibitors (Thermo Fisher Scientific, 2018):

i) if the VIC™ dye (IPC detector) amplified, but there is no amplification of the FAM™ dye (Quantifiler Human detector), this means that there is a successful PCR reaction (IPC target amplification), but there is no DNA in the sample, because amplification of FAM™ dye was unsuccessful. It is a true negative result;

ii) when the FAMTM dye amplified but there is no amplification of the VICTM dye, it is possible that the amount of DNA is very high (>10 ng/μl) or there are PCR inhibitors;

iii) if there is not any amplification of FAMTM dye or VICTM dye, it is a no valid result; possibly, there is no DNA in the sample or an unsuccessful reaction due to the presence of PCR inhibitors.

3.3. Bisulfite conversion

To evaluate DNAm levels across the selected DNA regions we used sodium bisulfite treatment introduced by Frommer *et al.* (1992), consisting in a chemical modification in which unmethylated cytosine residues are converted into uracil residues (U). Methylated cytosine (5mC or C^m) remains unchanged during bisulfite treatment (Grunau *et al.*, 2001; Genereux *et al.*, 2008) (**Figure 3.10**). The recovered DNA can be amplified by the polymerase chain reaction (PCR) following by Sanger sequencing, where U is detected as thymine (T).

Sodium bisulfite treatment requires a relatively low amount of DNA and can be applied to all CpG sites across the genome (Vidaki *et al.*, 2013), being very often accepted as the method of choice in forensic DNAm analysis.

In this work, genomic DNA from all tissue types (blood, bone, teeth and buccal swabs) was subjected to bisulfite conversion using the *EZ DNA Methylation-GoldTM Kit* (Zymo Research, Irvine, USA). Briefly, 20 μl of the extracted genomic DNA solution (in a total amount of 200 to 500 ng) was treated with sodium bisulfite and the modified DNA was extracted to a final volume of 10 μl. The procedure was performed according to the instructions of manufacturer, as follows:

i) preparing CT Conversion Reagent, and add 130 μl of CT Conversion Reagent to the 20 μl of genomic DNA (as the total amount of input DNA should be between 200-

500 ng for optimal results, some DNA samples were adjusted according DNA concentration);

ii) put the sample mixture in thermal cycler with the following conditions (98°C for 10 min; 64°C for 2.5 hours);

iii) put the sample into the Zymo-Spin™ IC Column with 600 µl of M-Binding Buffer, mixing by inverting the column several times;

iv) centrifugation at full speed (30 sec);

v) addition of 100 µl of M-Wash Buffer. Repeat the step iv;

vi) addition of 200 µl of M-Desulphonation Buffer, and incubate at room temperature for 15-20 min. Repeat the step iv;

vii) addition of 200 µl of M-Wash Buffer and centrifuge at full speed (30 sec).

Repeat this step;

viii) addition of 10 µl of M-Elution Buffer to the column matrix into a 1.5 ml microcentrifuge tube to elute DNA. Repeat the step iv;

ix) storage the DNA at or below -20°C; for long term storage, at or below -70°C.

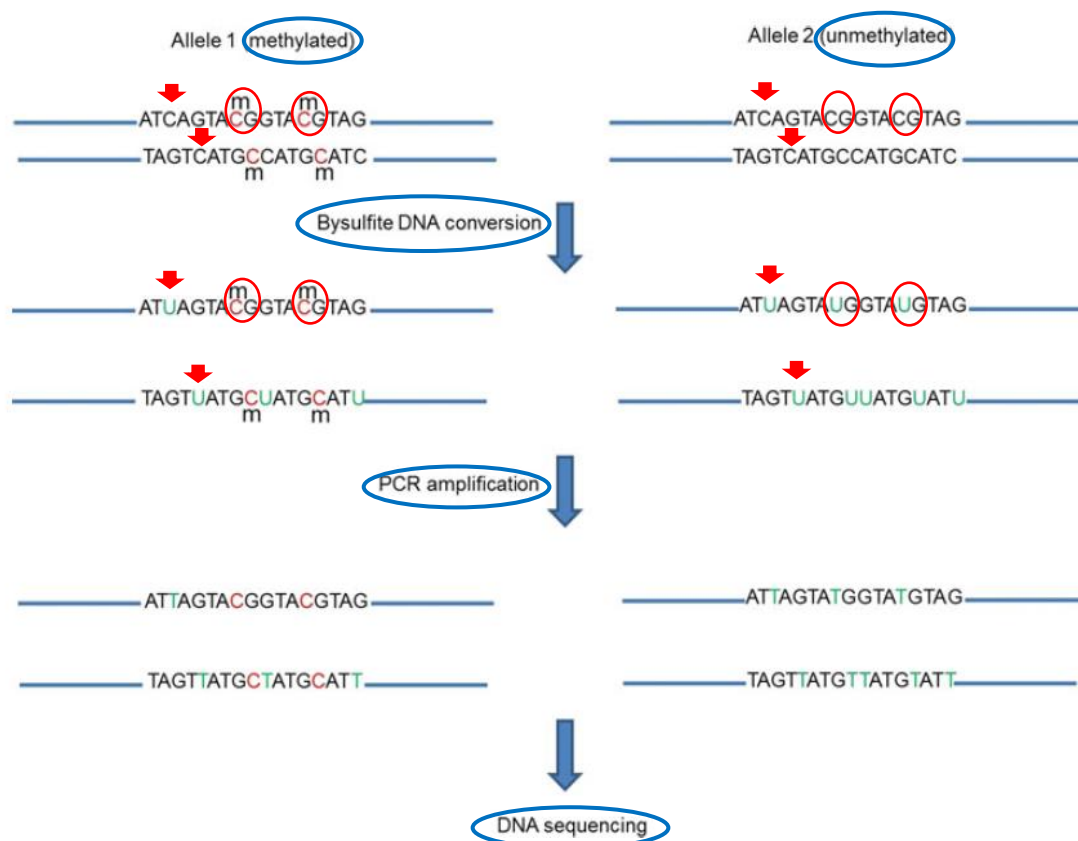


Figure 3.10: DNAm analysis based on bisulfite-PCR sequencing. After bisulfite conversion, in CpG sites (red circles) only the unmethylated cytosine is converted into U, while the 5mC or C^m remains as C. In non-CpG sites, the unmethylated cytosine is also transformed in U (red arrows). After PCR amplification all the U are displayed as T in the Sanger sequence of the sense strand. Adapted from GrŠković *et al.* (2013).

3.3.1. Efficiency of bisulfite conversion

The *EZ DNA Methylation-GoldTM Kit* allowed that almost all unmethylated cytosines (>99%) were converted to U, while more than 99% of 5mC is protected of conversion, according to the instructions of manufacturer (Zymo Research, Irvine, USA). This is important because bisulfite treatment requires the full conversion of unmethylated cytosine into U to obtain a correct analysis of DNAm levels at specific CpG sites.

After the bisulfite treatment, all the unmethylated cytosine residues in CpG and non-CpG positions were converted into U. As cytosines in non-CpGs are essentially unmethylated, in that positions all cytosines were converted into U. In this study, the efficiency of bisulfite conversion was estimated by measurement of the conversion of a

random number of cytosines at non-CpG positions. A mean conversion efficiency of 99.99% was observed over all samples and loci under study. This should mean that in almost all non-CpG sites, the conversion of C into U occurred completely and only the T was observed in sequencing chromatogram.

3.4. Polymerase Chain Reaction (PCR)

After bisulfite conversion, modified DNA samples were submitted to PCR reactions. The PCR amplification for selected regions of the target genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* was made using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany). Primer sequences were previously described in Bekaert *et al.* (2015a) and Zbieć-Piekarska *et al.* (2015b) (primer sequences are listed in **Table 3.2**).

A total PCR volume of 25 μ l was used consisting in a 1 μ l of the bisulfite converted DNA, 1 μ l (0.2 μ M final concentration) of each forward and reverse primer, 12.5 μ l of 2x Qiagen Multiplex PCR Master Mix (Qiagen) and 9.5 μ l of RNase-free water. PCR amplification was performed on a Biometra TProfessional thermocycler (Biometra GmbH) consisting in an initial step at 95°C for 15 min followed by 40 cycles of 30 sec at 95°C, 1 min at 60°C and 1 min at 72°C for the *ELOVL2* gene, and 35 cycles of 30 sec at 95°C, 1 min at 56°C and 1 min at 72°C for *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes. A final extension of 72°C for 10 min ended PCR amplification. A negative PCR control was included in each amplification. The size (addressed with GeneRuler Low Range DNA Ladder - Thermo Scientific) and quality of PCR products were visualized on 2% agarose gels. A strong band with the anticipated product size should be present for sequencing.

3.5. Sanger sequencing

Despite bisulfite Sanger sequencing (**Figure 3.11**) is not a new methodology in DNAm analysis, some reports addressed that it is very useful, rapid, cost-effective and highly accurate in detection of methylation levels of CpG sites (Jiang *et al.*, 2010; Parrish *et al.*, 2012).

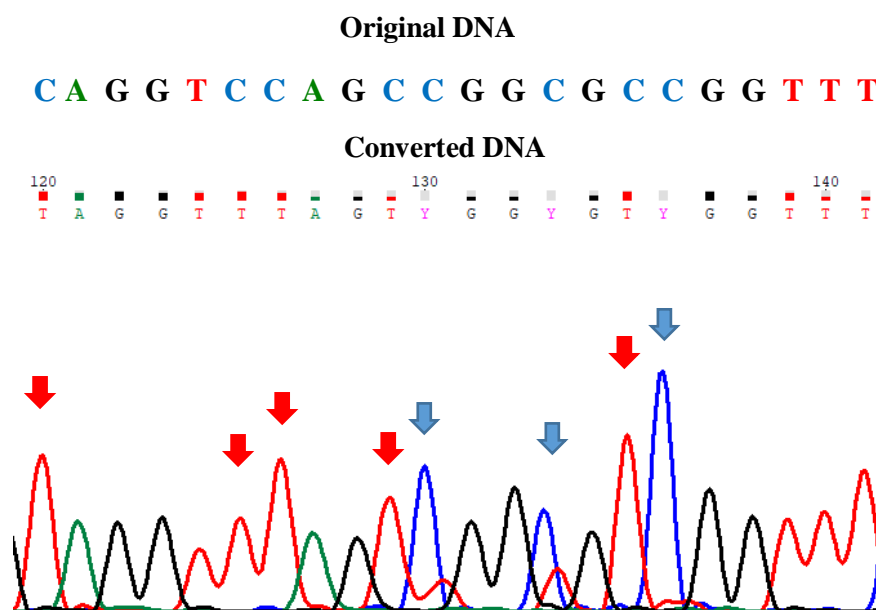


Figure 3.11: Sequencing data after bisulfite conversion and PCR amplification. Original DNA sequence reveals non-CpG sites and CpG sites. In chromatogram, red arrows represent converted cytosines at non-CpG sites and blue arrows show the CpG sites. Methylated cytosines at three CpG sites are represented by Y in the converted DNA sequence from the gene *ELOVL2*. G and A represent guanine and adenine, respectively.

An amount of 2 μ l of each PCR product was purified with 0.5 μ l of ExoSAP-IT (Affymetrix, Cleveland, USA) in a reaction with the following conditions: 37°C for 15 min, followed by 80°C for 15 min. The ExoSAP cleanup reagent removes the excess of primers and/or nucleotides not incorporated in the PCR reaction.

The Sanger dideoxy chain termination sequencing reaction was performed using, per sample: 2 μ l of the purified DNA, 0.5 μ l of reverse primer (100 ng/ μ l), 1 μ l of the reagent *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems, Foster City, USA) and 6.5 μ l of H₂O in a total volume of 10 μ l. The sequencing reaction was run in a

Biometra TProfessional thermocycler (Biometra GmbH) with the following conditions: initial denaturation at 96°C for 20 sec, following 25 cycles of denaturation at 96°C for 25 sec following annealing and extension at 60°C for 2 min.

The sequencing reactions were then purified using columns Optima™ DTR (Edge BioSystems, CA, USA) according to the manufacturer instructions, and applied directly on the ABI3130 genetic analyzer (Applied Biosystems, Foster City, USA) using the POP-7™ polymer as separation matrix. Sequenced samples were analyzed using the software Chromas (Version 2.32, Technelysium, Australia).

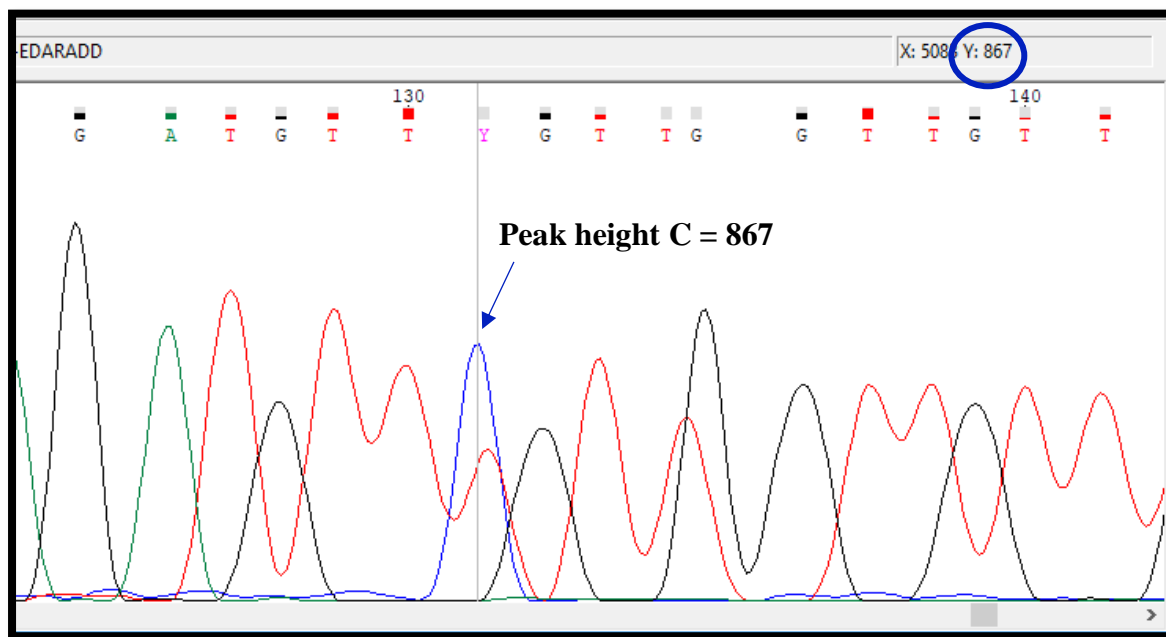
Table 3.2: PCR primers and sequence to analyse for PCR-sequencing.

Primers	Primer sequences	Reference
<i>ELOVL2</i> -for	AGGGGYGTAGGGTAAGTGAG	Bekaert <i>et al.</i> (2015a)
<i>ELOVL2</i> -rev	AAACCCAACATAAAACAAAACCAA	
Sequence to Analyse	TAGGTTTAGTYGGYGTGGTTTYGYGY GGYGGTTTAAAYGTTTAY	
<i>EDARADD</i> -for	TTGGTGATTAGGAGTTTTAGTGTTTT	Bekaert <i>et al.</i> (2015a)
<i>EDARADD</i> -rev	CCACCTACAAATCCCCAAA	
Sequence to Analyse	TTTAGTYGTTTTGAGGTTTATGGYGATG TTGAGTTTGGTTTTTAAATTTTTGGAGT TTGTTATGGAAGAAGTAATAGATTYGA GAAGATGTTYGTTGG	-
<i>FHL2</i> -for	TGTTTTTAGGGTTTTGGGAGTATAG	Zbieć-Piekarska <i>et al.</i> (2015b)
<i>FHL2</i> -rev	ACACCTCCTAAACTTCTCCAATCTCC	
Sequence to Analyse	AGTTATYGGGAGYGTGTTTTYGGYGT GGTTTTYGGGYGYGAGTTTTYGGAYGA GGTTTGGGYGYGGT	-
<i>PDE4C</i> -for	AGGTTTGTAGTAGGTTGAG	Bekaert <i>et al.</i> (2015a)
<i>PDE4C</i> -rev	AACTCAAATCCCTCTC	
Sequence to Analyse	TAGGTTGAGTYGTTTTYGYGGTYGTTAT AGTATGATTAGAGTTTTYGAAGTATTTGT GGYGGTAATTTYGGYGTTTTATTYGTAT TTAATAGYGTTTTTATTYGGATTYGATA	-
<i>C1orf132</i> -for	GTAAATATATAAGTGGGGGAAGAAGGG	Zbieć-Piekarska <i>et al.</i> (2015b)
<i>C1orf132</i> -rev	TTAATAAAACCAAATCTAAAACATTC	
Sequence to Analyse	AAGGGGGTTARGTTATTAAGTTTTGAAG TTRGTRGGATTATTTATRGTRGTTTGRGT AGATTT	-

3.5.1. Methylation quantification of CpGs after PCR direct DNA Sanger sequencing

The methylation status of C in each CpG dinucleotide was estimated according to Jiang *et al.* (2010) and Parrish *et al.* (2012), by measuring the peak height values of C and T through the formula $[C/C+T]$ in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium, Australia), as shown in **Figure 3.12**. In each CpG, a single C reveals complete methylation (100%), a single T shows complete unmethylation (0%) and overlapping C and T reveals partial methylation (0-100%). A number of examples for each gene were present in **Figure 3.13**.

As the reverse primer was used in the sequencing reaction, resulting in a cleaner chromatogram, the reverse-complement strand of the sequencing chromatogram was used to estimate the ratio between peak heights of C and T.



$$\% \text{DNAm} = \frac{\text{Peak height of C}}{\text{Peak height of C + T}} \times 100 \qquad \% \text{DNAm} = \frac{867}{(867 + 515)} \times 100 = 63\%$$

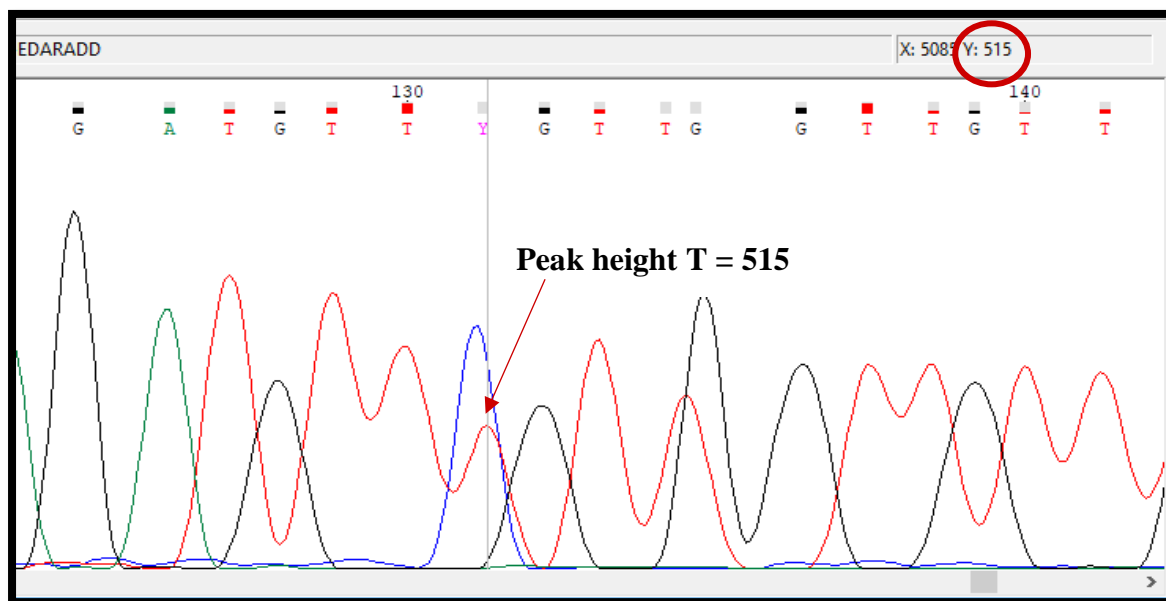
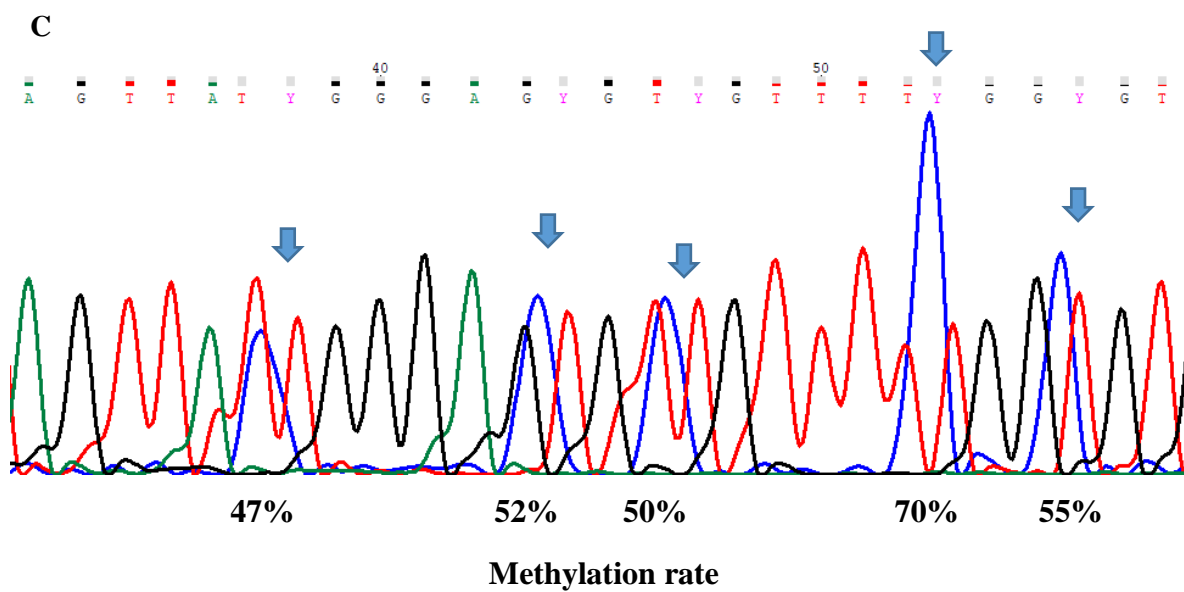
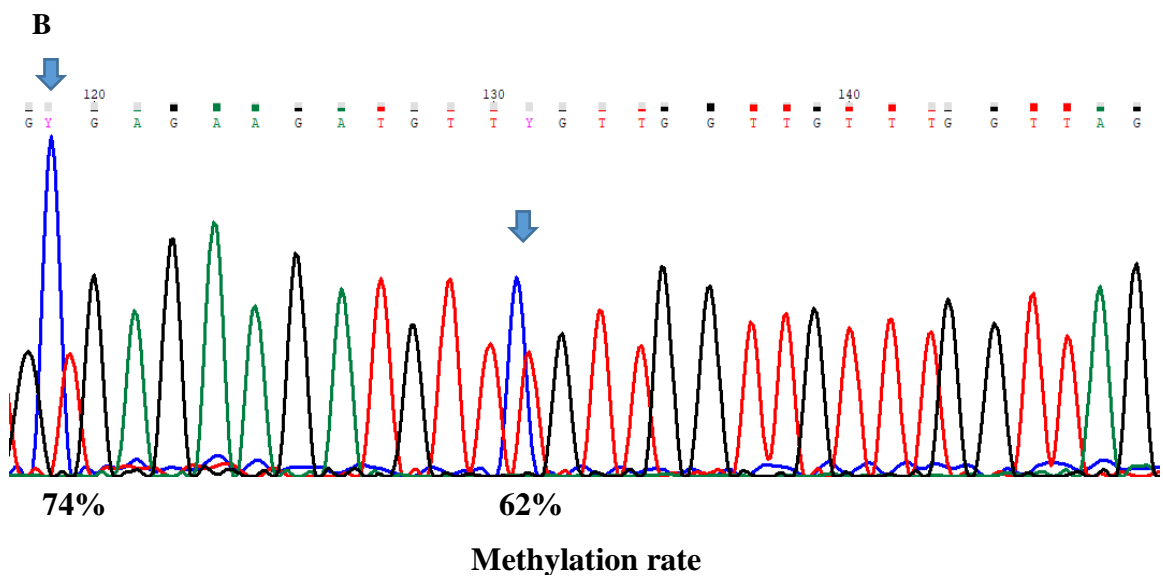
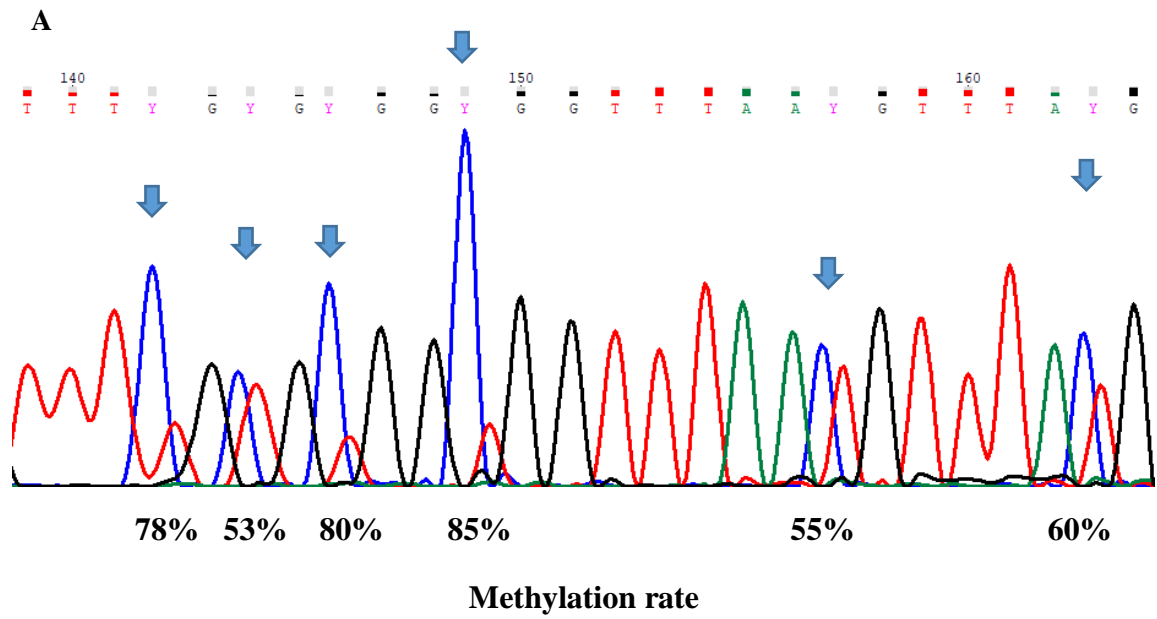


Figure 3.12: Peaks height determination in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium). The example shows the evaluation of DNAm levels of CpGs located at *EDARADD* locus in a younger adult individual (28 years, female). In each CpG, the blue arrow (peak height C) represents methylated cytosines and the red arrow (peak height T) represents unmethylated cytosines. The value of the blue and the red arrows are putted in the formula $[C/C+T]$ and the percentage of DNAm level is calculated.



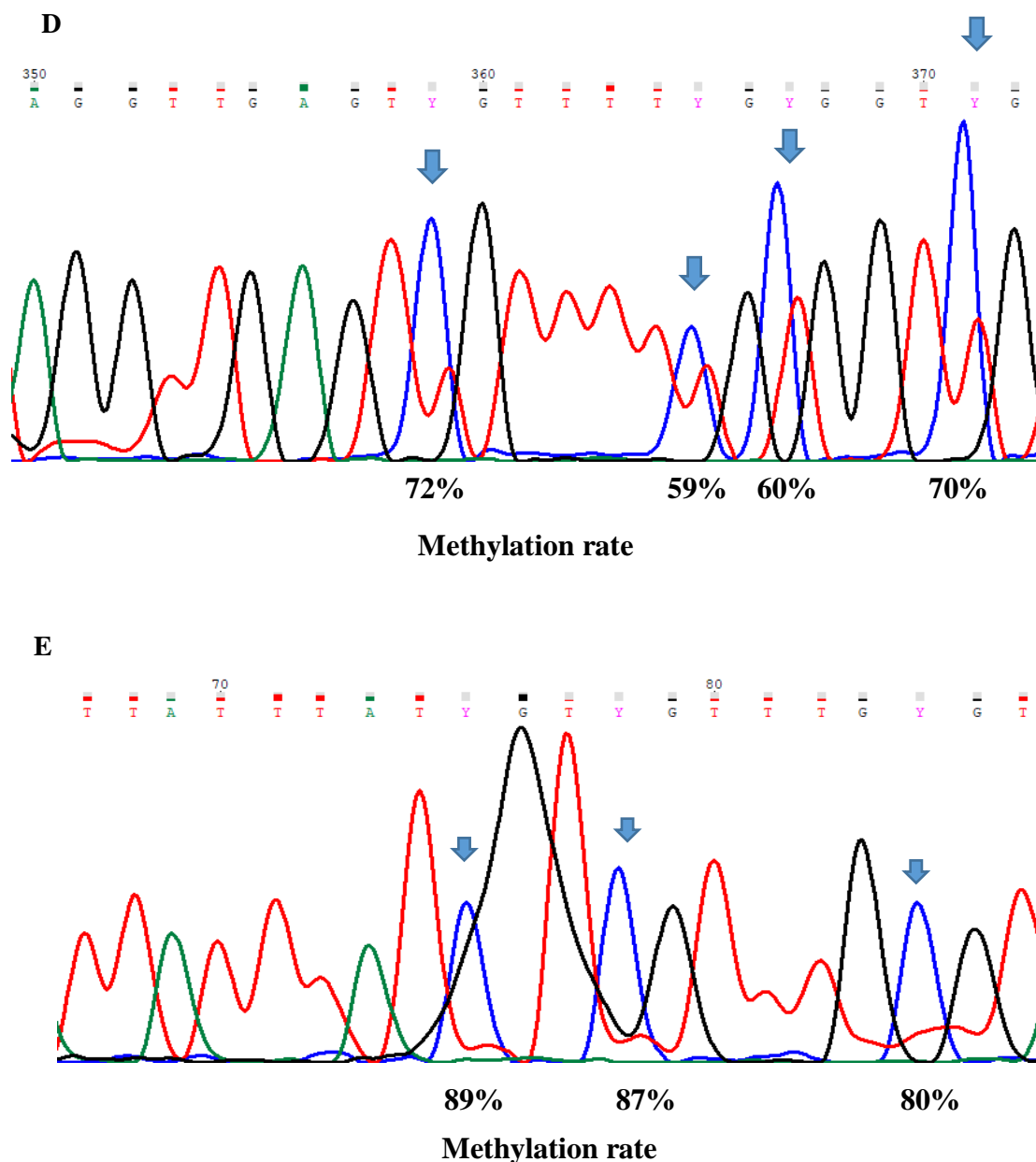


Figure 3.13: Examples of sequencing chromatograms for different methylation levels of the *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *Clorf132* genes by bisulfite PCR-sequencing. Blue arrows show CpG sites from each gene. A) Chromatogram of the *ELOVL2* gene for a blood sample from a living individual (37 years, female); B) Chromatogram of the *EDARADD* gene for a blood sample from a living individual (22 years, female); C) Chromatogram of the *FHL2* gene for a blood sample from a living individual (2 years, male); D) Chromatogram of the *PDE4C* gene for a blood sample from a living individual (94 years, female); E) Chromatogram of *Clorf132* gene for a blood sample from a deceased individual (35 years, male).

3.5.2. Reproducibility of direct bisulfite sequencing

To assess the reproducibility of the direct sequencing method, two separate PCR amplifications and Sanger sequencing analyses were performed in about 10% of the training set of blood samples for all genes. We checked if DNAm levels obtained in each CpG site under analysis are consistent between duplicate samples for the same individual.

3.5.3. DNAm standards

Each primer set used for bisulfite sequencing was independently verified, to confirm the accuracy of sequencing data using the DNAm commercial standards *EpiTect Control DNA*®, *methylated* and *EpiTect Control DNA*®, *unmethylated* (Qiagen, Hilden, Germany). Standard DNA samples premixed at methylation levels of 0%, 50%, and 100% were used for analysis. For each tissue type, the methylation levels obtained for the best-selected CpG sites were plotted against the expected DNAm levels.

3.6. Multiplex-PCR SNaPshot assay

After bisulfite conversion, modified DNA samples were analyzed for five CpG target sites located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* genes using the SNaPshot assay of Jung *et al.* (2019).

The capability of the SNaPshot® Multiplex System methodology allows to determine simultaneously DNAm levels of different CpGs from selected genes. The reaction is characterized by the multiplex amplification of several genomic regions and by a single-base extension (SBE) reaction, using multiplexed primers with variable size to target the different CpGs. The chemistry is based on the dideoxy SBE of unlabeled oligonucleotide primers that binds to the complementary template in the presence of fluorescently labeled ddNTPs and DNA polymerase (Applied Biosystems, 2005). The

polymerase extends the primer by one nucleotide, adding a single ddNTP to its 3' end (that represent the methylation signal of the targeted CpG site). The fluorescence color reports which base was added, that should be C (direct primer) or G (reverse primer) for the methylation signal or T (direct primer) or A (reverse primer) for the unmethylation signal (Applied Biosystems, 2005; Applied Biosystems, 2010).

The PCR amplification for the five CpG markers was performed using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and primers and conditions previously described in Jung *et al.* (2019), with some adaptations (primer sequences are listed in **Table 3.3**). The total PCR volume was 12.5 μ l, containing 1 μ l of converted DNA, 0.5 μ l of pool primers (μ M concentration of each primer is present in **Table 3.3**), 6.25 μ l of 2x Qiagen Multiplex PCR Master Mix (Qiagen) and 4.75 μ l of RNase-free water.

Table 3.3: Multiplex PCR primers adapted from Jung *et al.* (2019). The concentrations of *ELOVL2* and *TRIM59* genes were modified in the multiplex amplification SNaPshot reaction, to obtain a clear and strong band from these genes.

Genes (GRCh38)	Primer sequences (5'>3')	Concentration μM	Amplicon size (bp)
<i>ELOVL2</i> (Chr6:11044628)	F: GGGGYGTAGGGTAAGTGAG R: CAACRAATAAATATTCCTAAAACCTCC	0.6 0.6	187
<i>FHL2</i> (Chr2:105399282)	F: GGGTTTTGGGAGTATAGTAGT R: AAAATAACCCCTCCTCC	0.3 0.3	191
<i>KLF14</i> (Chr7:130734355)	F: AGGTTGTTGTAATTTAGAAGTTT R: ATATTTAACAACTCAAAAATTATCTTATC	0.3 0.3	114
<i>C1orf132</i> (Chr1:207823681)	F: GGGTTAYGTTATTAAGTTTTGAAG R: TAAAACCAAATTCTAAAACATTC	0.8 0.4	116
<i>TRIM59</i> (Chr3:160450189)	F: TATGGTATYGGTGGTTTGGGGGAGA R: ATAAAAAACACTACRCTCCACAACATAAC	0.4 0.4	148

Multiplex PCR amplification was performed in a Biometra TProfessional thermocycler (Biometra GmbH) and the program consisted in an initial step at 95°C for 11 min, followed by 25 cycles of 20 sec at 94°C, 1 min at 56 °C and 30 sec at 72°C. A final extension of 72°C for 7 min ended the multiplex PCR amplification. A negative PCR control (NTC) was included in each amplification. The size and quality of PCR products were visualized on 2% agarose gels with UV light (**Figure 3.14**).

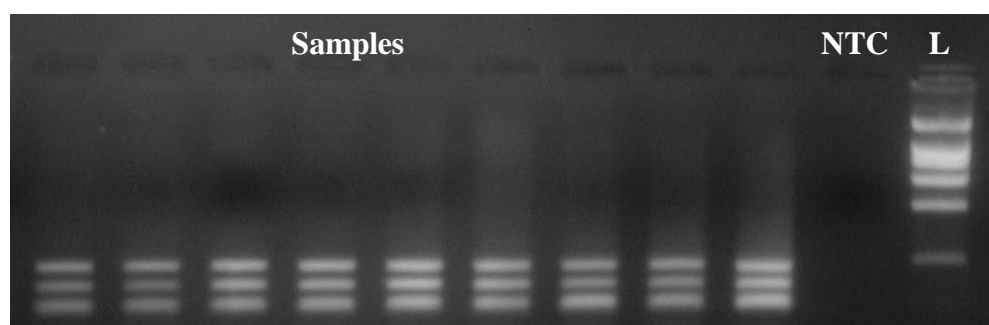


Figure 3.14: Agarose gel electrophoresis with the amplification products obtained using multiplex methylation SNaPshot assay for five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes. The first band represents the products of amplification of the *ELOVL2* (187bp) and *FHL2* (191bp), the second the amplification product of *TRIM59* (148bp) and the third band corresponds to the amplification products of *C1orf132* (116bp) and *KLF14* (114bp). NTC, negative PCR control; L, DNA Ladder.

Subsequently, 1 µl of the PCR products was purified with 0.5 µl of ExoSAP-IT (Affymetrix, Cleveland, USA) in an initial purification reaction (37°C for 15 min followed by 80°C for 15 min) (**Step 1, Figure 3.15**). After, the 1.5 µl of the purified PCR multiplex products was submitted to the SNaPshot reaction with 1 µl of SNaPshot Multiplex Kit (Applied Biosystems), 0.5 µl of pool primer mix (concentrations according **Table 3.4**) and 2 µl H₂O, in a total volume of 5 µl. Sequencing was performed in a Biometra TProfessional thermocycler (Biometra GmbH) with the following conditions: 10 sec at 96°C, following 5 sec at 50°C and 30 sec at 60°C, for 25 cycles (**Step 2, Figure 3.15**). A final purification was made using the 5 µl of the SNaPshot reaction with 1 µl of rSAP (Shrimp Alkaline Phosphatase Recombinant, Applied Biosystems) (37°C for 60

min and enzyme inactivation at 85°C 15 min) (**Step 2, Figure 3.15**). After that, 0.5 µl of each the purified SNaPshot reaction was diluted in 9.5 µl of a formamide solution with GS120 LIZ size standard (Applied Biosystems) (9 µl formamide + 0.5 µl GS120 LIZ), and the SBE reactions were run using the SeqStudio Genetic Analyzer (Applied Biosystems) and the GeneMapper Software 6 (Applied Biosystems) (**Step 3, Figure 3.15**). The methylation quantification was assessed in the electropherograms.

Table 3.4: Sequencing primers used for the multiplex methylation SNaPshot assay, according to Jung *et al.* (2019).

Genes	Primer sequences (5'>3')	Concentration (μM)	Length (nt)	Orientation
<i>ELOVL2</i>	(T) ₉ GGGAGGAGATTTGTAGGTTTAGT	5.0	32	F
<i>FHL2</i>	(T) ₂₁ GTTTTGGGAGTATAGTAGTTAT	0.5	43	F
<i>KLF14</i>	(T) ₂₈ TAAACAACCTCAAAAATTATCTTATCTCC	0.3	57	R
<i>C1orf132</i>	(T) ₄₆ AAACCAAAATTTAAATCTAC	1.2	66	R
<i>TRIM59</i>	(T) ₅₁ CCTCAAAAACCRTCRACCACCRAC	0.1	75	R

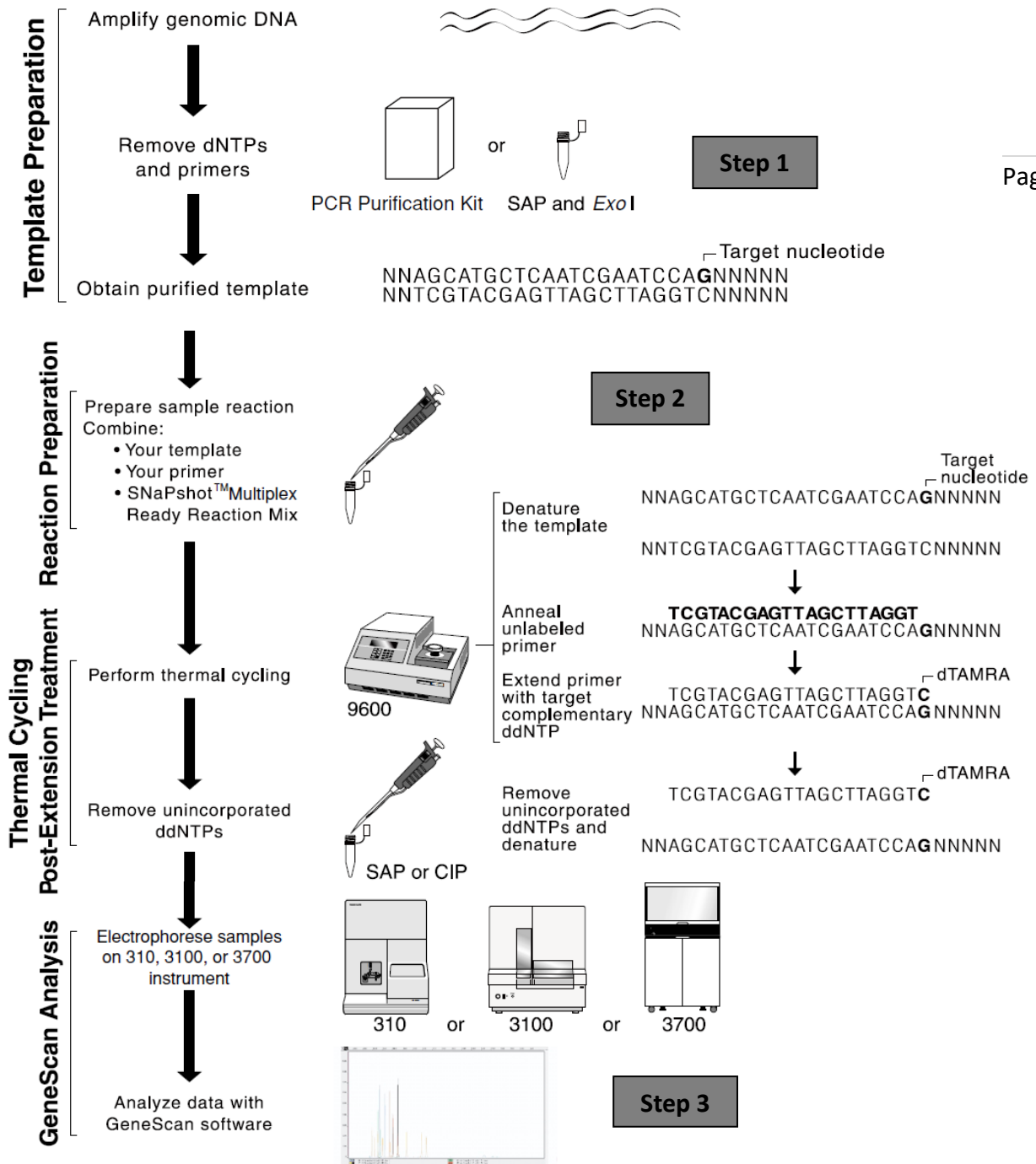


Figure 3.15: Methylation SNaPshot assay. Adapted from Quick Reference Card: ABI PRISM® SNaPshot™ Multiplex System (Applied Biosystems, 2010).

3.6.1. Methylation quantification of CpGs after multiplex SNaPshot assay

The methylation levels at each CpG site (0–1) from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were estimated from the nucleotide intensities measured by peak heights observed in the electropherograms, as described in Jung *et al.* (2019).

The methylation level from *ELOVL2* and *FHL2* CpGs was estimated using the formula $[C/C+T]$, in which C represents the methylation signal and T represents the non-methylation signal. The DNAm level from *KLF14*, *C1orf132* and *TRIM59* CpGs was measured by the equation $[G/G+A]$, in which the G is the reverse complement of C and the A is the reverse complement of T (**Figure 3.16**).

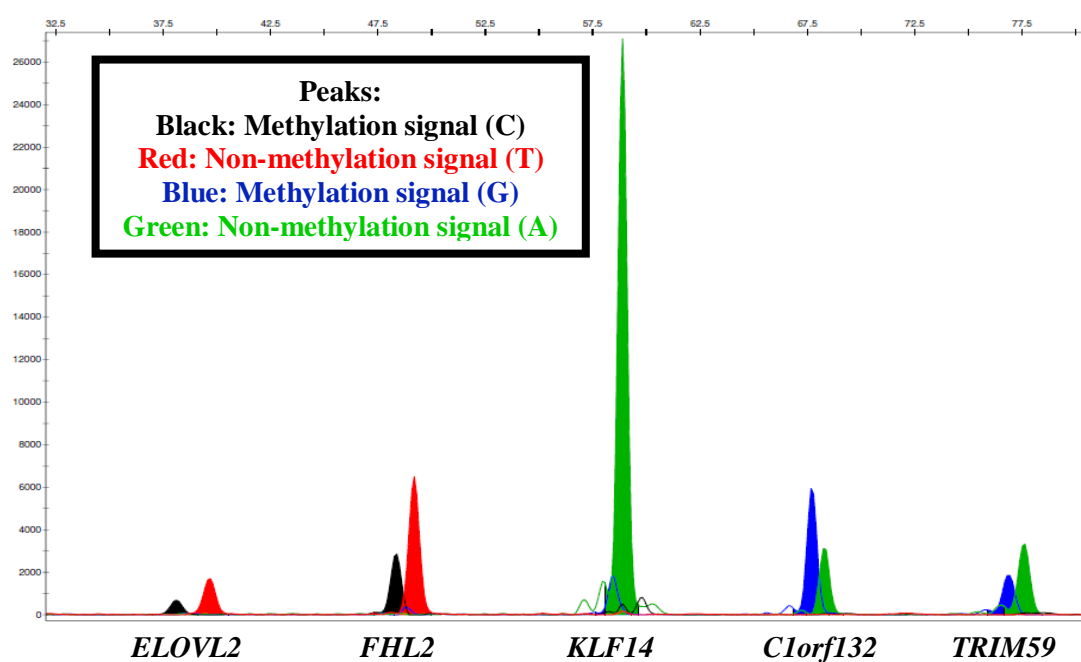


Figure 3.16: Electropherogram of the multiplex methylation SNaPshot assay for detection of DNAm levels in a blood sample from a female with 51 years old. Black (C) and blue (G) peaks represent methylation signal; red (T) and green (A) peaks represent non-methylation signal.

3.7. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Independent analyses were made for data obtained through Sanger sequencing and SNaPshot methodologies in each training set of different tissue types.

Linear regression models were used to analyze relationships between methylation levels of CpG sites and chronological age, calculating the coefficient of correlation R , R^2 , corrected R^2 and the p-value reflecting the statistical significance of the regression model. The unstandardized regression coefficients and corresponding p-values were also calculated to interpret the effect of each independent variable on the outcome. The normality of the dependent variable chronological age was assessed by Shapiro–Wilk and Kolmogorov–Smirnov tests. Multicollinearity was investigated by estimating Spearman correlation coefficients between predictor variables and the coefficient values $>70\%$ were considered as signal of multicollinearity. Cook’s distance was used to find influential outliers in all the predictor variables.

For Sanger sequencing data, using the simple linear regression coefficients from the highest age-correlated CpG sites from each gene, *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132*, we predicted age of individuals, in each training set or tissue combined training sets, according to the equation: $Y = b_0 + b_1x_1$, where: Y is the predicted age of the individual; b_0 is the y-intercept; b_1 is the slope of the line; and x_1 is the methylation value of the selected CpG (the independent variable). The highest age correlation CpG site (showing the highest value of the regression coefficient R) from each gene, or the best combination of CpG sites addressed by using a stepwise regression approach, were selected for simultaneous analysis using multiple linear regression to build a final age prediction model (APM). Using the multiple regression coefficients, we

predicted age of individuals applying the multiple linear regression formula: $Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Nx_N$, where: Y is the predicted age of the individual; b_0 is the y-intercept; b_1, b_2, b_N the slope of the selected CpGs; and x_1, x_2, x_N the methylation values of the selected CpGs.

For the SNaPshot methodology, simple linear regressions were used to analyze relationships between DNAm levels and chronological age at individual CpG sites located at each *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes. The simple linear regression coefficients of each significant age-correlated CpG site were used to predict age of individuals in each training set. Methylation information of the significant age-correlated CpGs was used in the stepwise regression analysis to select the best combination of the predictor variables to be used in the final multiple linear regression APMs.

Spearman correlation coefficients between predicted and chronological ages, the Mean Absolute Deviation (MAD) between chronological and predicted ages and the root mean square error (RMSE) were calculated for each training set of different tissue type or combination of tissues. For each training set the obtained MAD value was interpreted as either correct or incorrect if the predicted age was concordant with the chronological age using a cutoff value according to the standard error (SE) of estimate obtained in the developed APM. Moreover, MAD values and correct predictions were calculated for subsets of distinct age categories in the training sets of blood samples and buccal swabs.

Validation of the final APMs in each training set was performed by a K -fold cross validation that consists in removing randomly a set of samples from the training set and to develop K independent multiple linear regressions on the remaining samples. Subsequently, each APM is used to predict the age of the removed samples assigned as validation sets. The MAD values were obtained for each of the K independent multiple

linear regressions and the mean value was calculated. An additional validation was performed by splitting the complete data set into two subsets (training and validation sets) and an independent regression was calculated for the training set and applied to the validation set. In addition, when available an independent set of samples were used for validation purposes.

The assessment of differences in males and females was made using the comparison of regression lines using the STATGRAPICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA). For the evaluation of differences, we made a comparison of regression lines relating chronological age and DNAm levels of each gene at two levels (males / females) of the categorical factor. Analyses were made to determine if there are significant differences between the slopes and the intercepts at the two levels of that categorical factor.

The evaluation of the effect of PMI in DNAm levels was made considering six pairs of individuals from BDS with the same or similar chronological age (A, B, C, D, E and F). Individuals belonging to the pairs A, C and F have 1 year of difference between PMIs of the individuals of the pair, individuals of the pair D have 2 years of difference between the PMIs and individuals of the pairs B and E share the same PMI and were used as control.

Chapter 4. Results and discussion

Chapter 4. Results and discussion

A. DNA methylation age estimation in blood samples

The obtained data from blood samples from living and deceased individuals was used on published original papers:

Correia Dias H, Cordeiro C, Corte Real F, Cunha E, Manco L. Age estimation based on DNA methylation using blood samples from deceased individuals. Journal of Forensic Sciences 2019, 65(2): 465-470. DOI: 10.1111/1556-4029.14185.

Correia Dias H, Cordeiro C, Pereira J, Pinto C, Corte Real F, Cunha E, Manco L. DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay. Forensic Science International 2020 (311): 110267. DOI: 10.1016/j.forsciint.2020.110267.

Correia Dias H, Cunha E, Corte Real F, Manco L. Age prediction in living: Forensic epigenetic age estimation based on blood samples. Legal Medicine 2020 (47): 101763. DOI: 10.1016/j.legalmed.2020.101763.

1. Introduction

Several DNA methylation (DNAm) markers have been investigated in various tissues and body fluids using DNA-based methodologies such as bisulfite pyrosequencing (Weidner *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Eipel *et al.*, 2016; Cho *et al.*, 2017; Daunay *et al.*, 2019; Pfeifer *et al.*, 2020), EpiTYPER technology (Freire-Aradas *et al.*, 2016), massively parallel sequencing (Naue *et al.*, 2017, 2018; Aliferi *et al.*, 2018) or SNaPshot assays (Lee *et al.*, 2015; Hong *et al.*, 2017; Jung *et al.*, 2019). This allowed the identification of many CpG markers showing high correlations with chronological age, potential useful as forensic age predictors. Thus, a number of high accurate age prediction models (APMs) have been proposed for specific tissues, including blood (Zbieć-Piekarska *et al.*, 2015a, 2015b; Naue *et al.*, 2017), teeth (Giuliani *et al.*, 2016), buccal swabs (Eipel *et al.*, 2016) or saliva (Hong *et al.*, 2017), or as multi-tissue models (Horvath, 2013; Alsaleha *et al.*, 2017; Jung *et al.*, 2019).

Most of DNAm markers have been investigated and validated mainly in whole blood of living individuals using bisulfite pyrosequencing. Because DNAm at genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, *C1orf132*, *KLF14* and *TRIM59* have been repeatedly reported in independent studies to have strong age association in blood, they are considered to be some of the most promising age-predictive markers for blood. Both *ELOVL2* and *EDARADD* genes were included in a shortlist of 44 genomic regions most significantly associated with age in a meta-analysis (Bacalini *et al.*, 2017). Also, the *PDE4C* locus was ranked in the three best markers among 102 age-related CpG sites in blood (Weidner *et al.*, 2014). Zbieć-Piekarska *et al.* (2015b) using pyrosequencing published an assay which included five CpG sites in the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* genes, resulting in a Mean Absolute Deviation (MAD)

from the chronological age of 3.9 years in blood samples from 120 Polish individuals. Cho *et al.* (2017) replicated the strong age association for DNAm markers in the *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14* and *FHL2* genes using pyrosequencing methodology in blood samples of 100 Korean individuals obtaining a MAD of 4.2 years.

Meanwhile, the bisulfite polymerase chain reaction (PCR) sequencing methodology was shown to be an efficient and economical alternative tool for rapid quantification of DNAm, with similar linearity and accuracy than pyrosequencing analysis (Jiang *et al.*, 2010; Parrish *et al.*, 2012).

Recently, based on CpGs selected in the Zbieć-Piekarska *et al.* (2015b) model, Jung *et al.* (2019) tested the five CpG sites located in *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* genes for age prediction purposes in blood, saliva and buccal swab samples of healthy Korean individuals using the multiplex methylation SNaPshot method. The age-predictive linear regression model using the five CpG sites showed a high correlation between predicted and chronological ages in blood samples, with a MAD from chronological age of 3.174 years. Moreover, at least one or more of these genes were previously investigated for forensic purposes in blood samples of living (Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015a; Cho *et al.*, 2017) or deceased individuals (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Naue *et al.*, 2018; Pfeifer *et al.*, 2020).

Most studies on DNAm for age estimation purposes focused on the identification of new sets of markers in a specific population group or training specific methodologies. However, the validation and replication of experiments to test proposed age-predictive DNA markers and methodologies are strongly recommended for forensic applications to establish consistency between populations and laboratories (Cho *et al.*, 2017; Daunay *et al.*, 2019).

Moreover, the possibility that *postmortem* changes could alter the methylation status when performing age prediction should also be considered in forensic cases. To the best of our knowledge, only four studies have used blood samples from deceased individuals to address the correlation between DNAm and chronological age. The study of Bekaert *et al.* (2015a), using the pyrosequencing methodology, investigated CpG sites from four genes (*ASPA*, *PDE4C*, *ELOVL2* and *EDARADD*) in blood samples from 169 deceased and 37 living individuals. Naue *et al.* (2018) investigated 13 CpGs (located in genes *DDO*, *ELOVL2*, *F5*, *GRM2*, *HOXC4*, *KLF14*, *LDB2*, *MEIS1-AS3*, *NKIRAS2*, *RPA2*, *SAMD10*, *TRIM59* and *ZYG11A*) in blood samples from 29 deceased individuals by massively parallel sequencing. These markers were previously selected as strong age-dependent loci on whole blood from living individuals (Naue *et al.*, 2017). Hamano *et al.* (2016) analyzed the methylation levels of the *ELOVL2* and *FHL2* promoter regions by methylation-sensitive high resolution melting (MS-HRM) using 22 living and 52 dead blood samples. Finally, Pfeifer *et al.* (2020) evaluated DNAm levels of *PDE4C*, *ASPA*, *EDARADD* and *ELOVL2* genes by pyrosequencing in 151 blood samples from deceased individuals and compared with methylation information captured from 21 blood samples from living individuals.

That said, the main purposes of the current study were as follows:

- i) to develop specific APMs for blood samples from living and deceased individuals of Portuguese ancestry through the bisulfite PCR sequencing methodology using DNAm markers at genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132*;
- ii) to replicate the multiplex SNaPshot assay proposed by Jung *et al.* (2019) in blood samples from Portuguese individuals;
- iii) to compare DNAm status between different populations;

iv) to address putative differences in the methylation status between blood samples from living and deceased individuals.

2. Materials and Methods

2.1. Sample collection

Peripheral blood samples of 71 healthy individuals of Portuguese ancestry (45 females, 26 males; aged 1-95 years old), were collected in EDTA-tubes from users of *Biobanco - Hospital Pediátrico de Coimbra* and other hospitals for developing our DNAm model. Written informed consent was previously obtained from adult participants and from children's parents, under the age of 18 years.

A total of 73 blood samples (15 females, 58 males; aged 24-91 years old) was collected in EDTA-tubes during autopsy in *Serviço de Patologia Forense da Delegação do Centro do INMLCF*, after consulting RENNDA (*Registo Nacional de Não Dadores*) and from the Bodies Donated to Science (BDS) before the embalming method with Thiel (*Eisma et al., 2013*), in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* (Annex II). All blood samples from deceased individuals were collected within five days after death.

The study protocol was approved by the *Instituto Nacional de Medicina Legal e Ciências Forenses* and by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (n° 038-CE-2017).

2.2. DNA extraction, quantification and bisulfite conversion

Blood samples from living and deceased individuals were submitted to DNA extraction using the *QIAamp DNA Mini Kit* (Qiagen, Hilden, Germany), as previously

described in *Chapter 3. Sample and design research*. DNA extracts were quantified in a Nanodrop spectrophotometer (Thermo Fisher Scientific). Genomic DNA was subjected to bisulfite conversion using *EZ DNA Methylation-Gold™ Kit* (Zymo Research, Irvine, USA) according to the instructions of manufacturer (previously described in *Chapter 3. Sample and design research*). Briefly, 20 µl of genomic DNA (in a total amount of 200 to 500 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 µl.

2.3. Polymerase chain reaction (PCR) and Sanger sequencing

After bisulfite conversion, the modified DNA samples were submitted to PCR for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described in *Chapter 3. Sample and design research*. *C1orf132* gene was analyzed only for blood samples from deceased individuals. To assess the reproducibility of the method, two separate PCR amplifications and Sanger sequencing analyses were performed in about 10% of DNA samples for all genes.

Two independent investigators evaluated about 50% of the samples included in this study to indemnify the inter-observatory error.

2.4. SNaPshot assay

After bisulfite conversion, the modified DNA samples were submitted to a multiplex SNaPshot assay for the five CpG sites at genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* with the primers and conditions previously described in Jung *et al.* (2019). Particular conditions for multiplex PCR amplification and multiplex

SBE (single-base extension) reactions were as previously described in *Chapter 3. Sample and design research*.

2.5. DNAm quantification

Methylation quantification of C in each CpG dinucleotides evaluated by Sanger sequencing and SNaPshot methodologies was estimated according previously described in *Chapter 3. Sample and design research*.

2.6. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Linear regression models were used to analyze relationships between methylation levels and chronological age. Independent analyses were made for Sanger sequencing and SNaPshot methodologies in the set of blood from living individuals, deceased individuals and in the overall study sample (blood samples from living and deceased individuals).

For Sanger sequencing data, using the simple linear regression coefficients from the highest age-correlated CpG sites selected from each gene, we predicted age of individuals in blood samples from living and deceased individuals, and in the overall sample. The same highest age-correlated CpG sites in each set of samples were selected for simultaneous analysis using multiple linear regression to build specific final APMs for blood from living and deceased individuals and for the overall set of living and deceased individuals. A training set of 53 healthy individuals (35 females, 18 males; aged 1-95 years old) was used for development of the APM. In deceased individuals, the training set for development of APM consisted of 51 autopsies (7 females, 44 males; aged 24-86 years old).

For the multiplex SNaPshot assay, simple linear regression coefficients of each significant age-correlated CpG located at *ELOVL2*, *FHL2*, *KLF14*, *TRIM59* and *Clorf132* genes were used to predict age of individuals in blood samples from living, deceased and the overall set of samples. Methylation information of the significant age-correlated CpGs was used in the multiple regression approach to select the predictor variables to be used in the final multi-locus APMs. The training sets included peripheral blood samples from 59 Portuguese healthy living individuals (37 females, 22 males; aged 1-94 years old) and a total of 62 blood samples from deceased individuals (13 females, 49 males; aged 28-86 years old).

For both methodologies, the mean absolute deviation (MAD) between chronological and predicted ages and the root mean square error (RMSE) were calculated using the final APM for training sets of blood samples. MAD values were also calculated for subsets of distinct age categories in each training set. For living individuals the MAD values were calculated for subsets of four distinct age categories: <18 years old, 19-39 years old, 40-60 years old and >61 years old. For deceased individuals, three distinct age categories were addressed (24-51 years old, 52-71 years old and 72-88 years old). Each obtained result was interpreted as either correct or incorrect if the predicted age was concordant with the chronological age using a cutoff value according to the standard error (SE) of estimate obtained in the developed APM.

Validation of the final APMs developed for each set of samples and each methodology was performed by splitting the complete data set into two subsets (training and validation sets) and by a 4-fold cross validation as previously referred in *Chapter 3. Sample and design research*. Additionally, for the methylation information captured using bisulfite sequencing methodology an independent validation of the APM was performed using an independent set of 18 blood samples of healthy individuals (10

females, 8 males; aged 1-93 years old) and an independent set of 22 blood samples of deceased individuals (8 females, 14 males; aged 37-91 years old).

The evaluation of differences between sex was made through comparison of two regression lines relating chronological age and DNAm levels of each gene at two levels (males/females) of the categorical factor, using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA) in the training set of living individuals using both methodologies.

3. Results

In the present study we evaluated the DNAm levels of several CpG sites located at *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, *KLF14*, *C1orf132* and *TRIM59* genes through bisulfite conversion followed by PCR and direct Sanger sequencing or the SNaPshot methodology in blood samples from living and deceased individuals.

The efficiency of bisulfite conversion was confirmed measuring the conversion of a random number of cytosines at non-CpG positions. A mean conversion efficiency of 99.99% was observed over all samples and loci.

As Sanger sequencing is a semi-quantitative methodology for assessing DNAm levels, we tested the reproducibility of the method in about 10% of the blood samples of living and deceased individuals by two separate PCR amplifications and sequencing analyses for all genes. The obtained data revealed that DNAm levels obtained in each CpG site of *ELOVL2*, *EDARADD*, *PDE4C*, *C1orf132* and *FHL2* for the same individual are consistent in duplicate analysis (**Supplementary Figure S1**). The mean percentage difference in DNAm levels for all CpGs in each gene was 6.5% for *PDE4C* (two samples),

4.3% for *EDARADD* (seven samples), 2.9% for *ELOVL2* (five samples), 3.9% for *FHL2* (seven samples) and 7.6% for *C1orf132* (nine samples).

In addition, the accuracy of methylation levels obtained by bisulfite sequencing was evaluated by analyzing the PCR mixture amplification for each locus using three different methylation rates of 0%, 50%, and 100% (**Supplementary Figure S2**). For each best-selected site in blood samples from living and deceased individuals, bisulfite sequencing resulted in DNAm levels that bore a significant linear relationship to expected methylation levels (**Supplementary Figure S2**).

3.1. Age estimation in blood samples from living individuals

We evaluated DNAm levels of several CpG sites located at genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* through bisulfite conversion followed by PCR and direct Sanger sequencing in 71 blood samples from living Portuguese individuals. Moreover, we reanalyzed the multiplex methylation SNaPshot assay of Jung *et al.* (2019) testing the five CpG sites located in genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* in blood samples from 59 healthy Portuguese individuals.

3.1.1. DNAm data obtained in blood samples from living individuals using bisulfite Sanger sequencing

DNAm levels and sex

Two simple linear regression lines of methylation status and age between males and females showed no statistically significant difference in slope and intercept for target

sites in blood samples from living individuals (P -value >0.05 , **Table 4.1**). Thus, all the analyses were made ignoring differences due to sex.

Table 4.1: Comparison of two regression lines between males and females in blood samples from living individuals using data obtained from Sanger sequencing.

Marker	<i>P</i> -value	
	Intercept	Slope
<i>ELOVL2</i> CpG6	0.5884	0.7638
<i>FHL2</i> CpG3	0.9968	0.1959
<i>EDARADD</i> CpG3	0.3914	0.2013
<i>PDE4C</i> CpG2	0.2511	0.0612

Correlation between DNAm levels and chronological age

DNAm levels of 37 CpG sites located at *ELOVL2* (9 CpGs), *FHL2* (12 CpGs), *EDARADD* (4 CpGs) and *PDE4C* (12 CpGs) genes were analyzed in a training set of 53 blood samples from living Portuguese individuals (35 females, 18 males; aged 1-95 years old). Positive correlations between methylation levels and chronological age were observed for all examined CpG sites of the *ELOVL2* gene and negative correlations were observed for CpG2, CpG3 and CpG4 at *EDARADD* gene (**Supplementary Table S1**). For *FHL2* and *PDE4C* genes, different correlations were observed among the tested CpG markers: clear positive correlations were observed for *FHL2* CpG1 to CpG6 and CpG8, and *PDE4C* CpG1 to CpG5; and clear negative correlations were observed for *FHL2* CpG11, CpG12, and *PDE4C* CpG6, CpG9 (**Supplementary Table S1**). **Supplementary Figure S3** shows the correlation between age and DNAm levels of the best site in these four genes.

Development of an age prediction model (APM)

Simple linear regressions testing the correlation between methylation levels and chronological age, revealed significant associations (P -value <0.05) for all CpGs from the genes *ELOVL2* and *FHL2* (except CpG7, CpG9 and CpG10) (**Supplementary Table S1**). Significant associations with methylation levels were also observed for *EDARADD* CpG2, CpG3 and CpG4, as well as for *PDE4C* CpG1 to CpG6 and CpG9 (**Supplementary Table S1**).

The *FHL2* CpG3 ($R = 0.940$, P -value = 1.78×10^{-25}) was the most powerful CpG site considering all CpGs analyzed in the four genes, explaining 88.1% of the variation in age, followed by *ELOVL2* CpG6 ($R = 0.936$; P -value = 7.97×10^{-25}), explaining 87.4% of the variation in age, *EDARADD* CpG3 ($R = -0.888$; P -value = 7.86×10^{-19}), explaining 78.4% of the variation in age, and *PDE4C* CpG2 ($R = 0.852$; P -value = 6.32×10^{-16}), explaining 72% of the variation in age (**Table 4.2; Supplementary Table S1**). The predicted age of individuals was calculated through the simple linear regression coefficients for the individual strongest age-associated markers and the obtained MAD values between predicted and chronological ages were as follows: 8.01 years for *ELOVL2* CpG6; 7.81 years for *FHL2* CpG3; 10.57 years for *EDARADD* CpG3; and 11.87 years for *PDE4C* CpG2 (**Table 4.2; Supplementary Figure S4**).

Table 4.2: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in blood samples from living individuals.

Locus	CpG site	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	CpG6	Chr6:11044644	53	0.936	0.877	0.874	10.41	7.97×10^{-25}	8.01
<i>FHL2</i>	CpG3	Chr2:105399291	53	0.940	0.884	0.881	10.11	1.78×10^{-25}	7.81
<i>EDARADD</i>	CpG3	Chr1:236394382	53	-0.888	0.788	0.784	13.63	7.86×10^{-19}	10.57
<i>PDE4C</i>	CpG2	Chr19:18233133	53	0.852	0.725	0.720	15.53	6.32×10^{-16}	11.87
<i>Multiple linear regression</i>									
APM (<i>ELOVL2</i> CpG6, <i>FHL2</i> CpG3, <i>EDARADD</i> CpG3 and <i>PDE4C</i> CpG2)			53	0.972	0.946	0.941	7.13	1.11×10^{-29}	5.35

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Combining the methylation information of CpGs most highly associated with age per locus, *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2, the multi-locus APM reveals a higher age correlation value ($R = 0.972$), highly significant (P -value = 1.11×10^{-29}), explaining 94.1% of the variation in age (corrected $R^2 = 0.941$) (**Table 4.2**). The value of corrected R^2 demonstrated that only around 6% of the variation in age cannot be explained by that model (**Table 4.2**). All the selected predictor variables showed a significant p -value justifying their inclusion in the final model (**Table 4.3**).

Table 4.3: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* in blood samples from living individuals.

Marker	Coefficient	<i>P</i> -value
(Intercept)	-81.879	0.011
<i>ELOVL2</i> CpG6	103.031	0.000
<i>FHL2</i> CpG3	99.331	0.001
<i>EDARADD</i> CpG3	-58.970	0.028
<i>PDE4C</i> CpG2	35.843	0.013

Predicted age of individuals was calculated through the multiple linear regression coefficients (**Table 4.3**) using the formula: $(-81.879) + 103.031 \times \text{DNAm level } ELOVL2 \text{ CpG6} + 99.331 \times \text{DNAm level } FHL2 \text{ CpG3} - 58.97 \times \text{DNAm level } EDARADD \text{ CpG3} + 35.843 \times \text{DNAm level } PDE4C \text{ CpG2}$. A strong correlation between predicted and chronological ages was obtained (Spearman correlation coefficient, $r = 0.972$) with a MAD from chronological age of 5.35 years (RMSE = 6.64) (**Figure 4.1**). Correct predictions were 75.5% assuming that chronological and predicted ages match ± 7 years, according to the standard error of estimate calculated for the final APM (SE = 7.13) (**Table 4.2**).

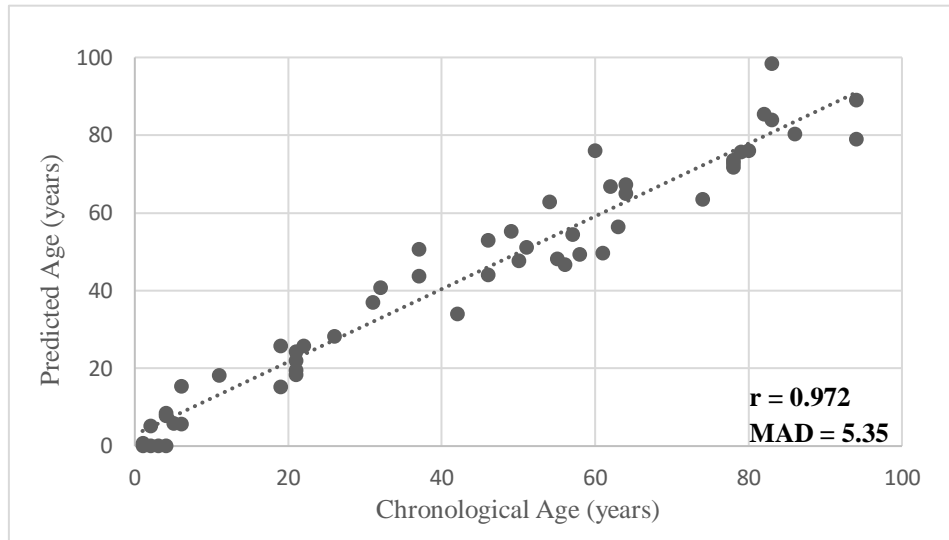


Figure 4.1: Predicted age *versus* chronological age using the four best markers *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 in blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Evaluating the differences between ages using the multiple APM developed in living individuals, larger differences between predicted and chronological ages were observed with increasing age (**Figure 4.2**).

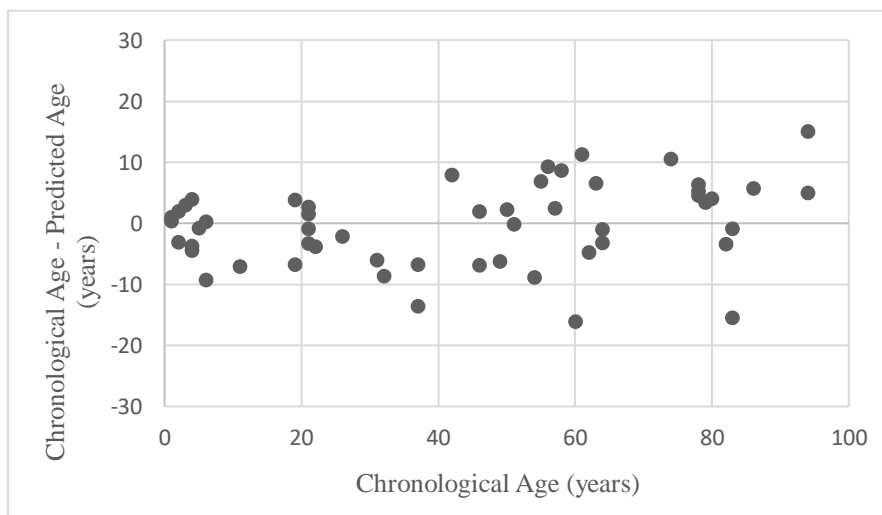


Figure 4.2: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from living individuals.

To investigate these age-related differences, we divided our training set of 53 living individuals in four age groups (<18 years old; 18-39 years old; 40-60 years old; >61 years old) to estimate MAD and percentage of correct predictions in each age range group (**Table 4.4; Figure 4.3**). Age predictions were considered either correct or incorrect if the predicted age was concordant with the chronological age ± 7 years, according to the standard error of estimate calculated for the final APM (SE = 7.13).

The MAD value is higher in the two older age categories (G3 and G4 age range groups): 40-60 years old (MAD = 6.49 years) and >61 years old (MAD = 6.27 years). In concordance, the lower percentage of correct predictions was observed in the older age groups G3 and G4 (58.3% and 76.5%, respectively). For younger individuals <18 years old (G1) and age range group 19-39 years old (G2), the smaller MAD values were obtained (MAD = 3.26 and 4.99 years, respectively) and the higher values of correct predictions were observed (83.3% in both age groups) (**Table 4.4; Figure 4.3**).

Table 4.4: MAD between predicted and chronological ages stratified by age group in the training set of 53 blood samples from living individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<18 years	12	3.26	83.3
G2	19-39 years	12	4.99	83.3
G3	40-60 years	12	6.49	58.3
G4	>61 years	17	6.27	76.5
Overall	1-95 years	53	5.35	75.5

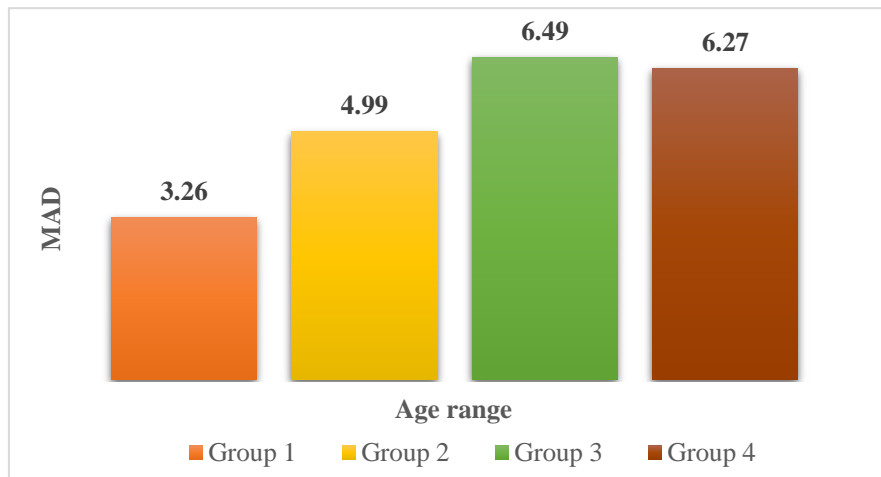


Figure 4.3: MAD from chronological age calculated for each age group in blood samples from living individuals. The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in living individuals

For evaluation of the accuracy of our multi-locus APM with *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 markers we made a 4-fold cross validation using our data set of 53 blood samples from living individuals. The mean of MAD values obtained amongst the four validation sets was 6.20 years (RMSE = 6.27), similar to the obtained in the overall population (MAD = 5.35 years).

The validation approach through splitting the overall sample set of 53 living into two sets of 27 and 26 samples (training and validation sets) allowed to obtain a MAD value of 6.08 years (RMSE = 7.45) in the new training set of 27 samples. Applying this multiple model on the validation set of 26 samples, a MAD of 5.81 years was obtained (RMSE = 6.51). Both independent MAD values were very close to the MAD of 5.35 years for the overall sample.

Additionally, using an independent sample set of 18 blood samples from healthy Portuguese individuals (10 females, 8 males; aged 1-93 years old), we evaluated the performance of our multi-locus APM developed in the overall sample of 53 living individuals. Based on the multiple linear regression model a strong correlation between

predicted and chronological ages was observed (Spearman correlation coefficient, $r = 0.977$), with a MAD from the chronological age of 4.98 years (RMSE = 6.58) (**Figure 4.4**).

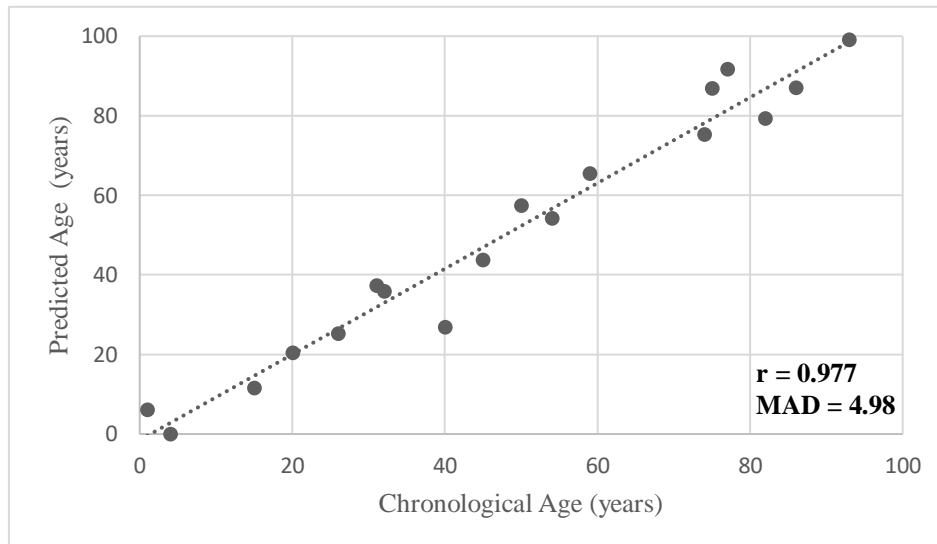


Figure 4.4: Predicted age *versus* chronological age of the test set, 18 living individuals, using the final model developed for blood samples from living individuals with the markers *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2. MAD value and Spearman correlation coefficient, r , are plotted on the chart.

In spite of the smallest number of samples in the test set of 18 individuals, evaluating the differences according to age ranges, we observed also an increase of MAD values with the increase of age (**Table 4.5; Figure 4.5**). Age groups <18 and 19–39 were grouped together because of the low number of samples.

Table 4.5: MAD between predicted and chronological ages stratified by age group in the test set of 18 blood samples from living individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1 + G2	0-39 years	7	3.38	100
G3	40-60 years	5	5.70	60.0
G4	>61 years	6	6.25	66.7
Overall	1-95 years	18	4.98	77.8

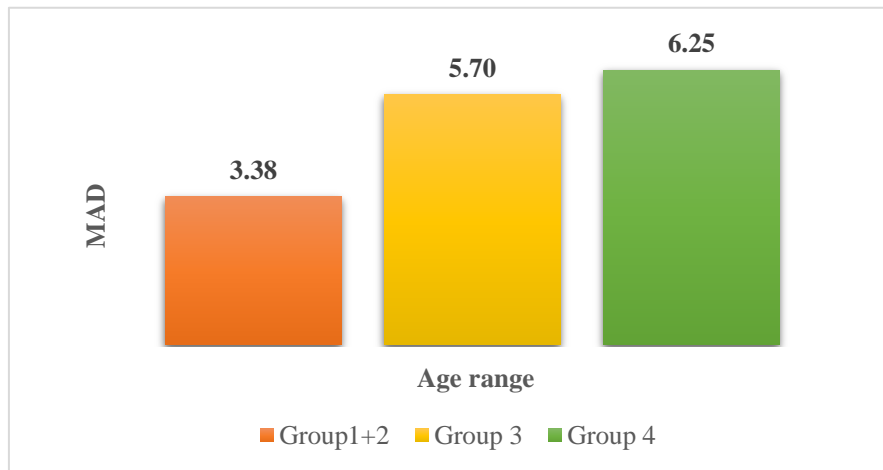


Figure 4.5: MAD from chronological age in the test set, 18 samples, using the developed APM for blood samples from living individuals. The MAD values are printed on the top of each age range. Age groups 1 and 2 were grouped together.

3.1.2. DNAm data obtained in blood samples from living individuals using SNaPshot methodology

Correlation between DNAm levels and chronological age

DNAm levels at the five CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were simultaneously measured through a SNaPshot assay in peripheral blood samples from 59 healthy Portuguese individuals (22 males, 37 females; aged 1-94 years old). Three samples did not amplify for *ELOVL2* gene and one sample did not amplify for *KLF14* gene.

Two simple linear regression lines of methylation status and age between males and females showed no statistically significant difference in slope and intercept in the sample set of living individuals (except in slope for *C1orf132*) (**Table 4.6**). Thus, all the analyses were made ignoring sex differences.

Table 4.6: Comparison of two regression lines between males and females in blood samples from living individuals using data obtained from SNaPshot methodology.

Marker	P-value	
	Intercept	Slope
<i>ELOVL2</i>	0.4473	0.9326
<i>FHL2</i>	0.2844	0.7228
<i>TRIM59</i>	0.2984	0.2098
<i>KLF14</i>	0.4401	0.3436
<i>C1orf132</i>	0.8803	0.0080

Positive correlations between DNAm and chronological age were observed for *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes and a negative correlation with age was obtained for *C1orf132* gene (**Supplementary Figure S5**).

Development of an age prediction model (APM)

Using simple linear regression, all the five CpG sites showed strong correlations between DNAm and chronological age (**Supplementary Figure S5**). The strongest correlation was observed for *ELOVL2* gene ($R = 0.951$, $P\text{-value} = 3.58 \times 10^{-29}$), explaining 90.2% of the variation in age, followed by *FHL2* ($R = 0.946$, $P\text{-value} = 1.49 \times 10^{-29}$), explaining 89.3% of the variation in age, *C1orf132* ($R = -0.924$, $P\text{-value} = 1.67 \times 10^{-25}$), explaining 85.2% of the variation in age, and *TRIM59* ($R = 0.910$, $P\text{-value} = 2.04 \times 10^{-23}$), explaining 82.4% of the variation in age. The *KLF14* gene showed the lowest age correlation ($R = 0.791$, $P\text{-value} = 1.57 \times 10^{-13}$), explaining 61.8% of the variation in age (**Table 4.7**). Predicting age through the simple linear regression equation for each CpG site revealed MAD values of 6.73 years for *ELOVL2*, 7.40 years for *FHL2*, 8.29 years for *TRIM59*, 8.80 years for *C1orf132*, and 13.46 years for *KLF14* (**Table 4.7**; **Supplementary Figure S6**).

Table 4.7: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using a SNaPshot assay in 59 blood samples from living individuals.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>ELOVL2</i>	Chr6:11044628	56	0.951	0.904	0.902	9.18	3.58×10^{-29}	6.73
<i>FHL2</i>	Chr2:105399282	59	0.946	0.895	0.893	9.42	1.49×10^{-29}	7.40
<i>C1orf132</i>	Chr1:207823681	59	-0.924	0.854	0.852	11.09	1.67×10^{-25}	8.80
<i>TRIM59</i>	Chr3:160450189	59	0.910	0.828	0.824	12.07	2.04×10^{-23}	8.29
<i>KLF14</i>	Chr7:130734355	58	0.791	0.625	0.618	17.59	1.57×10^{-13}	13.46
<i>Multiple linear regression</i>								
APM (<i>ELOVL2</i> , <i>FHL2</i> and <i>C1orf132</i>)		56	0.982	0.965	0.963	5.65	7.315×10^{-38}	4.25

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

We tested the age-predictive multiple linear regression model using simultaneously all the five CpG sites in the training set of 59 blood samples of living individuals. Although individually all the five CpG sites showed strong and significant associations with age (**Table 4.7**), in the multivariate analysis the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values (P -value = 0.479 and P -value = 0.948, respectively), which could reveal signs of multicollinearity between variables (**Supplementary Table S2**). In concordance, when we applied the stepwise linear regression analysis, the same three significant CpG sites were chosen (located at *ELOVL2*, *FHL2* and *C1orf132* genes) (**Table 4.8**). Thus, we built a final model with these three significant CpGs which reveals a higher age correlation value ($R = 0.982$), explaining 96.3% of the variation in age (corrected $R^2 = 0.963$), highly significant (P -value = 7.315×10^{-38}) (**Table 4.7**). Applying this model, the age prediction for each individual was obtained through the formula (**Table 4.8**): $38.751 + 61.058 \times \text{DNAm level } ELOVL2 + 80.021 \times \text{DNAm level } FHL2 - 47.631 \times \text{DNAm level } C1orf132$.

Table 4.8: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *FHL2* and *C1orf132*, selected by stepwise regression approach, in blood samples from living individuals.

Marker	Coefficient	P -value
(Intercept)	38.751	0.000
<i>ELOVL2</i>	61.058	0.000
<i>FHL2</i>	80.021	0.000
<i>C1orf132</i>	-47.631	0.000

The developed APM enabled us to estimate age with a correlation between predicted and chronological ages of 0.969 (Spearman correlation coefficient, $r = 0.969$) and a MAD from chronological age of 4.25 years (RMSE = 5.39) (**Figure 4.6**). Correct

predictions were 75% assuming that chronological and predicted ages match ± 6 years, according to the standard error of estimate calculated for the final APM (SE = 5.65).

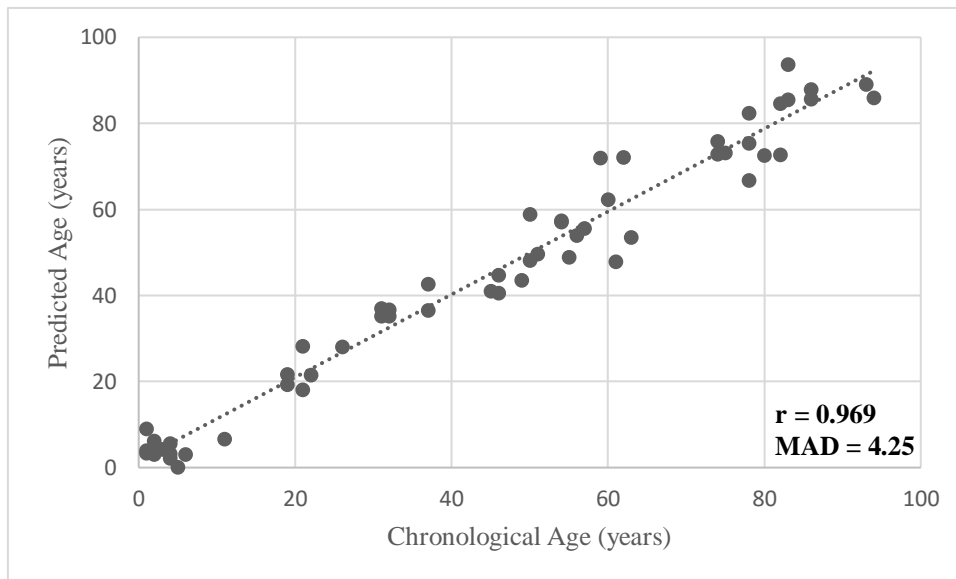


Figure 4.6: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay developed with the three CpGs located at *ELOVL2*, *FHL2* and *C1orf132* genes in blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Evaluating the differences between ages using the final APM developed in living individuals with the three CpGs located at *ELOVL2*, *FHL2* and *C1orf132*, larger differences between predicted and chronological ages were observed with increasing age (**Figure 4.7**).

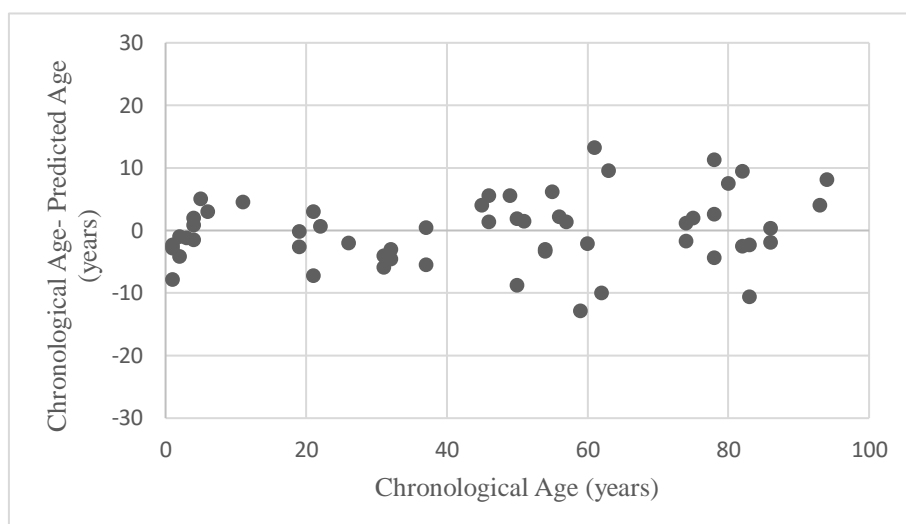


Figure 4.7: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from living individuals.

The MAD values between predicted and chronological ages increase with the increasing of age in blood samples of living individuals: the highest MAD value was obtained for the age group >61 years old (MAD = 5.70 years) and the smallest for age group <18 years old (MAD = 3.03 years) (**Table 4.9; Figure 4.8**). In concordance, the percentage of correct predictions were higher in younger age groups (**Table 4.9**).

Table 4.9: MAD between predicted and chronological ages stratified by age group in the training set of 56 blood samples from living individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<18 years	12	3.03	91.7
G2	19-39 years	12	3.28	83.4
G3	40-60 years	14	4.26	78.6
G4	>61 years	18	5.70	55.6
Overall	1-94 years	56	4.25	75.0

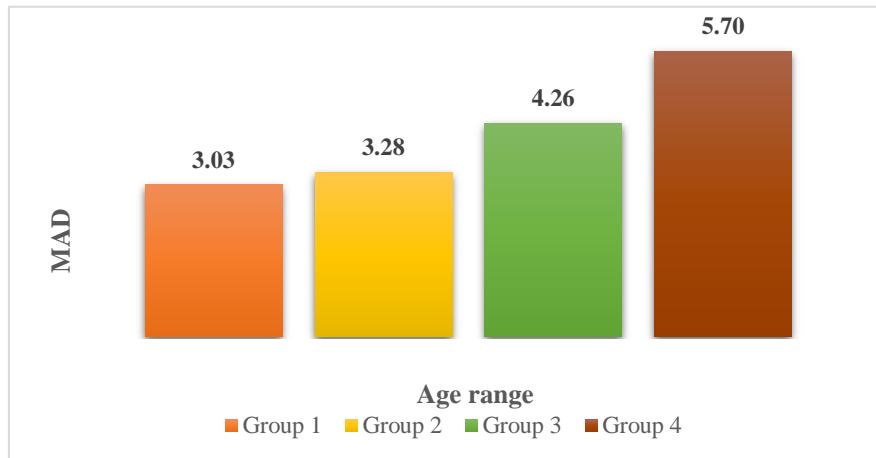


Figure 4.8: MAD from chronological age calculated for each age group in blood samples from living individuals. The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in living individuals

The accuracy of the final APM built with CpGs located at *ELOVL2*, *FHL2* and *C1orf132* was tested by a 4-fold cross validation. The MAD value between predicted and chronological ages was obtained in the validation set for each of the four independent multiple linear regressions and the calculated mean value reveals an average MAD value from chronological age of 4.75 years (RMSE = 4.77) amongst the four test sets, very close to the MAD of 4.25 years from the whole training set.

As an additional validation, we split the dataset into two similar sets of 27 and 29 samples each (training and validation set) and re-fitted the multivariate linear regression model on the training set. This allowed us to obtain an independent MAD value for the training set of 3.86 years (RMSE = 5.09). When this model was applied to the validation set, a MAD of 4.93 years was obtained (RMSE = 6.37). Both independent MAD values were very close to the MAD of 4.25 obtained from the whole data set (56 individuals).

3.1.3. Comparison between methodologies

Comparing age correlation values of the two CpGs (*ELOVL2*, Chr6:11044628 and *FHL2*, Chr2:105399282) analyzed by both bisulfite sequencing and SNaPshot methodologies in the set of living individuals, we observed similar and very strong age correlation values (*ELOVL2*, R = 0.920 vs. R = 0.951 and *FHL2*, R = 0.937 vs. R = 0.946, respectively) (Table 4.10).

Table 4.10: Comparison of age-correlated values obtained in blood samples from living Portuguese individuals through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Blood from living (1-95 years)		Blood from living (1-94 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.920	0.846	0.951	0.904
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.937	0.878	0.946	0.895

Abbreviations: R, Pearson correlation coefficient.

3.2. Age estimation in blood samples from deceased individuals

In the present section, we evaluated the association between chronological age and DNAm levels at genes *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132* in blood samples from deceased individuals, using the bisulfite PCR sequencing methodology. Moreover, we analyzed DNAm levels at the five CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes in blood samples of deceased individuals, through the multiplex methylation SNaPshot assay described in Jung *et al.* (2019).

As we only have 15 females in our data set of blood samples from deceased individuals evaluated by the Sanger sequencing methodology (7 females in training set,

8 females in an independent test set) and 13 females in the training set evaluated by the SNaPshot method, no statistically analysis was made for comparison between males and females. Hence, all the analyses were made ignoring sex differences.

3.2.1. DNAm data obtained in blood samples from deceased individuals using bisulfite Sanger sequencing

Correlation between DNAm levels and chronological age

Forty-three CpG sites (*ELOVL2*: 9 CpGs; *EDARADD*: 4 CpGs; *FHL2*: 12 CpGs; *PDE4C*: 12 CpGs; and *C1orf132*: 6 CpGs) were selected for methylation evaluation through the bisulfite PCR sequencing method in the training set of 51 blood samples (7 females, 44 males; aged 24-86 years old) from deceased individuals. The focused CpG positions were those selected in previous studies, with some adjacent CpG sites also considered in the analysis. For the set of 51 blood from deceased individuals, two samples did not amplify for *C1orf132* gene and two samples did not amplify for *ELOVL2* and *PDE4C* genes.

To obtain a first overview of the markers and the change of DNAm per year, a linear regression for individual CpG sites was performed in the training set of 51 individuals, after addressing the normal distribution of the dependent variable (**Supplementary Table S3**). The *ELOVL2* locus showed highly significant values for all the selected CpG sites ($R \geq 0.66$) reflecting the similar strength of the change in DNAm with age across all CpGs. For the remaining four loci, several CpG sites revealed no age-dependency (non-significant p-values) and some CpGs, although significantly associated, revealed moderate or low correlation values. For *FHL2*, all selected CpGs showed lower or moderate change of DNAm with age ($R < 0.50$) (**Supplementary Table S3**).

Development of an age prediction model (APM)

Simple and multiple APMs were developed for the most significant age-associated CpG marker in each locus, which mean the CpG with the higher age correlation coefficient and more significant p-value. A clear positive correlation between DNAm levels and age was observed for *ELOVL2* CpG4, *PDE4C* CpG2 and *FHL2* CpG2 markers and a clear negative correlation was observed for *C1orf132* CpG1 and *EDARADD* CpG3 markers (**Supplementary Figure S7**).

The *ELOVL2* CpG4 showed the strongest correlation with age ($R = 0.785$; P -value = 2.39×10^{-11}), explaining 60.8% of the variation in age, followed by *C1orf132* CpG1 ($R = -0.634$; P -value = 0.000001), explaining 38.9% of the variation in age. For the remaining genes, the strongest age-correlated sites were: *PDE4C* CpG2 ($R = 0.592$; P -value = 0.000008), explaining 33.6% of the variation in age; *EDARADD* CpG3 ($R = -0.621$; P -value = 0.000001), explaining 37.3% of the variation in age; and *FHL2* CpG2 ($R = 0.465$; P -value = 0.00058), explaining 20% of the variation in age (**Table 4.11; Supplementary Table S3**).

Predicting age based on DNAm levels of the best sites through the simple linear regression coefficients allowed to obtain a MAD of 8.89 years for *ELOVL2* CpG4, 9.35 years for *PDE4C* CpG2, 9.38 years for *C1orf132* CpG1, 9.93 years for *EDARADD* CpG3 and 11.40 years for *FHL2* CpG2 (**Table 4.11; Supplementary Figure S8**).

Because DNAm for the best CpG markers showed a linear relationship with age, no statistical transformation of variables was made. In spite of this, as a few existing studies refer a quadratic regression for *ELOVL2* (Bekaert *et al.*, 2015a) we tested the better fit for the relationship between DNAm levels and chronological age. The linear regression model showed the best result for *ELOVL2* CpG4 (linear: $R = 0.785$, $MAD = 8.89$, $RMSE = 9.99$; quadratic: $R = 0.780$, $MAD = 8.48$, $RMSE = 10.10$).

Table 4.11: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between DNAm levels obtained by bisulfite sequencing and chronological age in blood samples from deceased individuals.

Locus	CpG site	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	CpG4	Chr6:11044640	49	0.785	0.617	0.608	10.20	2.39×10^{-11}	8.89
<i>C1orf132</i>	CpG1	Chr1:207823681	49	-0.634	0.402	0.389	12.68	0.000001	9.38
<i>EDARADD</i>	CpG3	Chr1:236394382	51	-0.621	0.385	0.373	12.72	0.000001	9.93
<i>PDE4C</i>	CpG2	Chr19:18233133	49	0.592	0.350	0.336	13.29	0.000008	9.35
<i>FHL2</i>	CpG2	Chr2:105399288	51	0.465	0.216	0.200	14.36	0.000584	11.40
<i>Multiple linear regression</i>									
APM (<i>ELOVL2</i> CpG4, <i>C1orf132</i> CpG1, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2 and <i>FHL2</i> CpG2)			47	0.888	0.788	0.763	8.02	8.17×10^{-13}	6.08

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

After addressing the absence of multicollinearity between the five most significant age-associated CpG sites (*ELOVL2* CpG4, *C1orf132* CpG1, *EDARADD* CpG3, *PDE4C* CpG2, *FHL2* CpG2), and confirming the absence of outliers (Cook's $D < 1$), we tested a stepwise model by multiple linear regression joining successively the five markers (data not shown). An increase on age-associated statistical values was observed with the successive addition of CpGs. The overall APM using the five CpGs showed the most high correlation coefficient ($R = 0.888$), highly significant ($P\text{-value} = 8.17 \times 10^{-13}$), explaining 76.3% of the variation in age (**Table 4.11**). Applying this model, the age prediction for each individual was obtained with the multiple linear regression coefficients (**Table 4.12**) through the formula: $20.495 + 77.938 \times \text{DNAm level } ELOVL2 \text{ CpG4} + 46.879 \times \text{DNAm level } FHL2 \text{ CpG2} - 70.729 \times \text{DNAm level } EDARADD \text{ CpG3} + 28.741 \times \text{DNAm level } PDE4C \text{ CpG2} - 47.984 \times \text{DNAm level } C1orf132 \text{ CpG1}$. **Figure 4.9** presents a plot with chronological age *versus* predicted age in the training set of 47 individuals. A strong correlation between predicted and chronological ages was observed (Spearman correlation coefficient, $r = 0.868$), with a MAD of 6.08 years (**Figure 4.9**). The success rate of correct predictions was 63.8% assuming that chronological and predicted ages match ± 8 years (according to the SE of 8.02, **Table 4.11**).

Table 4.12: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* in blood samples from deceased individuals.

Marker	Coefficient	P-value
(Intercept)	20.495	0.559
<i>ELOVL2</i> CpG4	77.938	0.001
<i>C1orf132</i> CpG1	-47.984	0.015
<i>EDARADD</i> CpG3	-70.729	0.001
<i>PDE4C</i> CpG2	28.741	0.196
<i>FHL2</i> CpG2	46.879	0.030

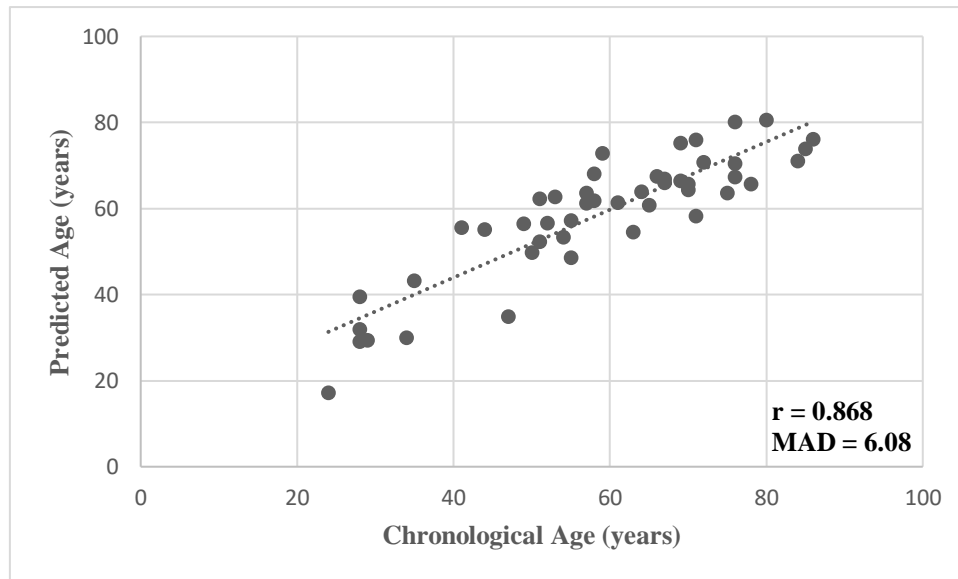


Figure 4.9: Predicted age *versus* chronological age using the five best markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *PDE4C* CpG2 and *Clorf132* CpG1 in blood samples from deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Evaluating the differences between ages using the multi-locus APM developed for blood samples from deceased individuals through bisulfite PCR sequencing we observed some differences between age categories (**Figure 4.10**).

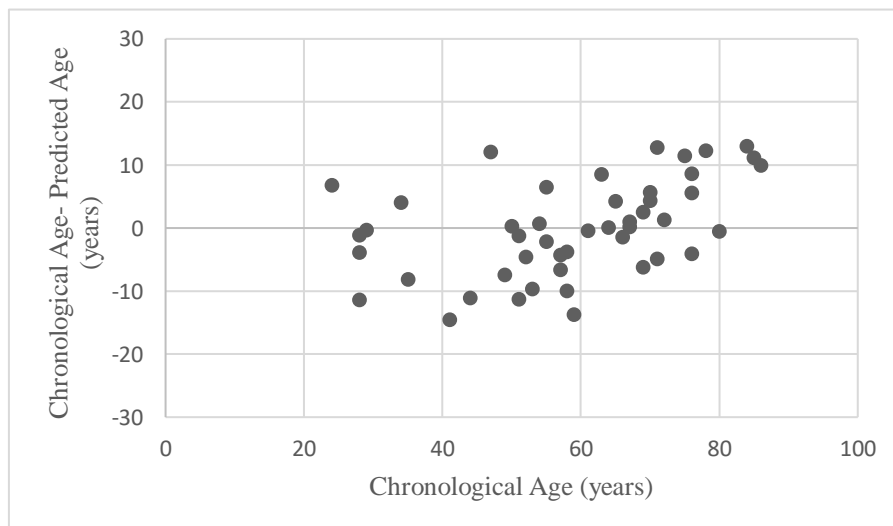


Figure 4.10: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from deceased individuals.

To investigate these age-related differences, we divided our training sample set in three age groups (Group 1: 24-51 years old; Group 2: 52-71 years old; Group 3: 72-86 years old), and then we calculated the MAD values and percentage of correct predictions (Table 4.13; Figure 4.11). A higher MAD value was observed in G3, which is the group that included older age categories (72-86 years old). Also, the percentage of correct predictions showed to be the lowest in this age range (Table 4.13).

Table 4.13: MAD between predicted and chronological ages stratified by age group in the training set of 47 blood samples from deceased individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	24-51 years	14	6.69	57.1
G2	52-71 years	23	4.97	78.3
G3	72-86 years	10	7.79	40.0
Overall	24-86 years	47	6.08	63.8

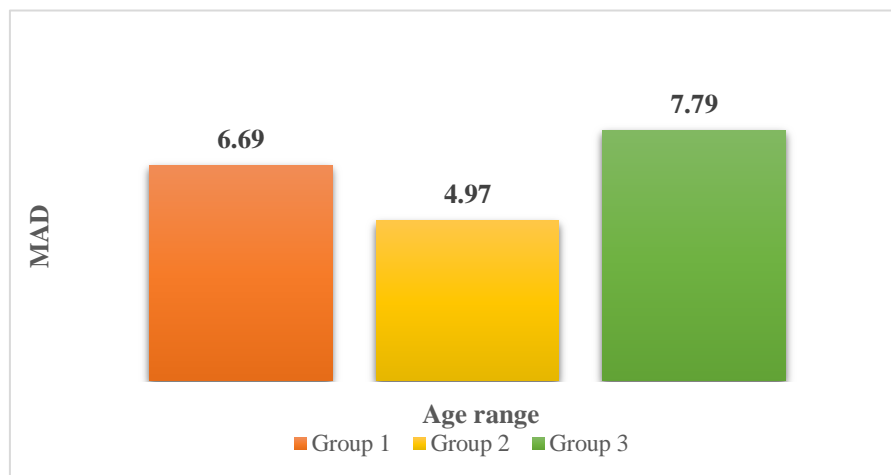


Figure 4.11: MAD from chronological age calculated for each age group in blood samples from deceased individuals. The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in deceased individuals

The model accuracy of the APM, based on DNAm levels of *ELOVL2* CpG4, *C1orf132* CpG1, *FHL2* CpG2, *PDE4C* CpG2 and *EDARADD* CpG3 markers, was

evaluated through a 4-fold cross validation in the training set of 47 deceased individuals, producing a MAD (mean value obtained for the four test sets) of 7.22 years (RMSE = 7.43). This value was very close to the MAD of 6.08 (RMSE = 7.49) obtained in the whole training set.

The validation by splitting the overall training set into two sets of 24 and 23 samples (training and validation sets) allowed to obtain an independent MAD value for the training set of 6.11 years (RMSE = 7.38). Applying the model on the validation set, a MAD of 7.23 years (RMSE = 8.35) was obtained. Both independent MAD values were very close to the MAD of 6.08 years (RMSE = 7.49).

Additionally, an independent set of 22 blood samples from deceased individuals (8 females, 14 males; aged 37-91 years old) was used to test the accuracy of the developed APM. In validation set, three samples did not amplify for *C1orf132* and one of these three samples did not amplify for *EDARADD*, *ELOVL2* and *PDE4C*. The validation in 19 samples revealed a MAD of 8.84 years (RMSE = 10.98) (**Figure 4.12**).

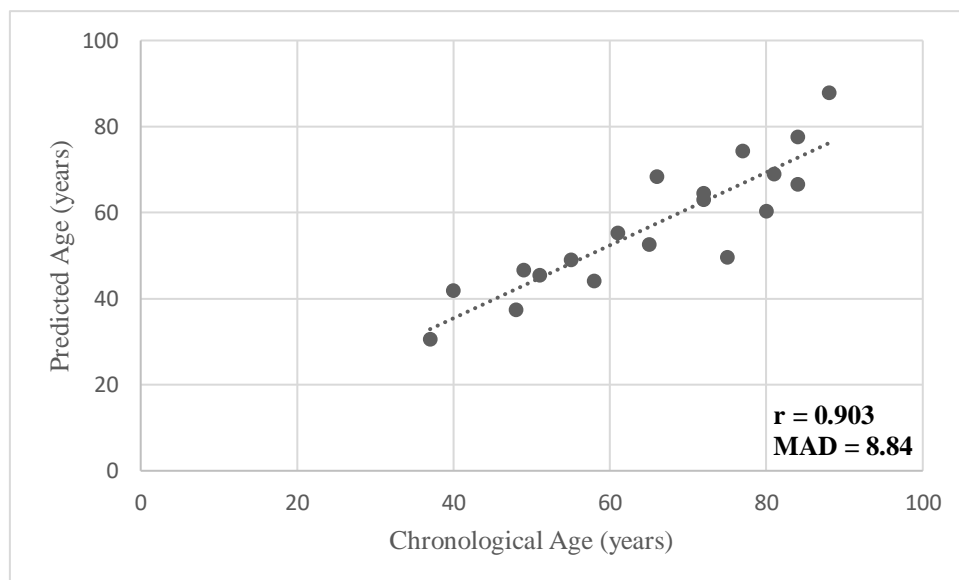


Figure 4.12: Predicted age *versus* chronological age of the test set, 19 deceased individuals, using the final best model developed for blood samples from deceased individuals including the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *C1orf132* CpG1 and *PDE4C* CpG2. MAD and Spearman correlation coefficient, r , are plotted on the chart.

3.2.2. DNAm data obtained in blood samples from deceased individuals using SNaPshot methodology

Correlation between DNAm levels and chronological age

A total of 62 blood samples from deceased individuals (49 males, 13 females; aged 28-88 years old) was tested using a multiplex methylation SNaPshot assay described in Jung *et al.* (2019). Some samples did not amplify: one sample for *TRIM59* gene, three samples for *KLF14* gene and two of these three samples also did not amplify for *C1orf132* gene. Among the five markers simultaneously analyzed using the multiplex SNaPshot assay, positive correlations were observed for *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes and a negative correlation was obtained for *C1orf132* gene (**Supplementary Figure S9**).

Development of an age prediction model (APM)

Testing age association among the five CpGs through simple linear regression, the CpG site located at *ELOVL2* showed the strongest correlation between DNAm and chronological age ($R = 0.791$, $P\text{-value} = 2.04 \times 10^{-14}$), explaining 61.9% of the variation in age, followed by *TRIM59* ($R = 0.769$, $P\text{-value} = 4.78 \times 10^{-13}$), explaining 58.4% of the variation in age, *FHL2* ($R = 0.654$, $P\text{-value} = 8.16 \times 10^{-9}$), explaining 41.8% of the variation in age, *C1orf132* ($R = -0.591$, $P\text{-value} = 6.56 \times 10^{-7}$), explaining 33.8% of the variation in age, and finally *KLF14* ($R = 0.568$, $P\text{-value} = 0.000003$), explaining 33.1% of the variation in age (**Table 4.14**). Simple APMs were developed for each CpG site and the obtained MAD values from chronological age were 7.64 years, 7.97 years, 9.81 years, 9.99 years and 10.40 years for *ELOVL2*, *TRIM59*, *FHL2*, *C1orf132* and *KLF14* genes, respectively (**Table 4.14; Supplementary Figure S10**).

Table 4.14: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot assay in 62 blood samples from deceased individuals.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>ELOVL2</i>	Chr6:11044628	62	0.791	0.626	0.619	9.75	2.04×10^{-14}	7.64
<i>TRIM59</i>	Chr3:160450189	61	0.769	0.591	0.584	10.24	4.78×10^{-13}	7.97
<i>FHL2</i>	Chr2:105399282	62	0.654	0.428	0.418	12.06	8.16×10^{-9}	9.81
<i>C1orf132</i>	Chr1:207823681	60	-0.591	0.350	0.338	12.93	6.56×10^{-7}	9.99
<i>KLF14</i>	Chr7:130734355	59	0.568	0.322	0.311	12.86	0.000003	10.40
<i>Multiple linear regression</i>								
APM (<i>ELOVL2</i> , <i>FHL2</i> , <i>C1orf132</i> and <i>TRIM59</i>)		59	0.899	0.808	0.793	7.26	1.07×10^{-18}	5.36

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

We tested the age-prediction multiple linear regression model using simultaneously these CpG sites in the training set of 59 blood samples from deceased individuals, however the CpG site at *KLF14* gene revealed a non-significant age correlation value (P -value = 0.424; **Supplementary Table S4**).

Furthermore, we used the stepwise regression to select the best model and the four CpG sites at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes were chosen. The final model constructed with these sites showed high age correlation coefficients ($R = 0.899$), explaining 79.3% of the variation in age (corrected $R^2 = 0.793$), highly significant (P -value = 1.07×10^{-18}) (**Table 4.14**). The developed formula obtained with regression coefficients to calculate age was the following (**Table 4.15**): $14.914 + 63.627 \times \text{DNAm level } ELOVL2 + 40.299 \times \text{DNAm level } FHL2 - 24.185 \times \text{DNAm level } C1orf132 + 57.717 \times \text{DNAm level } TRIM59$. The model showed a strong correlation between predicted and chronological ages (Spearman correlation coefficient, $r = 0.916$), with a MAD of 5.36 years (RMSE = 6.94) (**Figure 4.13**). Correct predictions were 72.9% assuming the standard error of estimate calculated for the final APM (SE = 7.26) (**Table 4.14**).

Table 4.15: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*, selected by stepwise regression approach, in blood samples from deceased individuals.

Marker	Coefficient	P -value
(Intercept)	14.914	0.043
<i>ELOVL2</i>	63.627	0.007
<i>FHL2</i>	40.299	0.019
<i>C1orf132</i>	-24.185	0.000
<i>TRIM59</i>	57.717	0.001

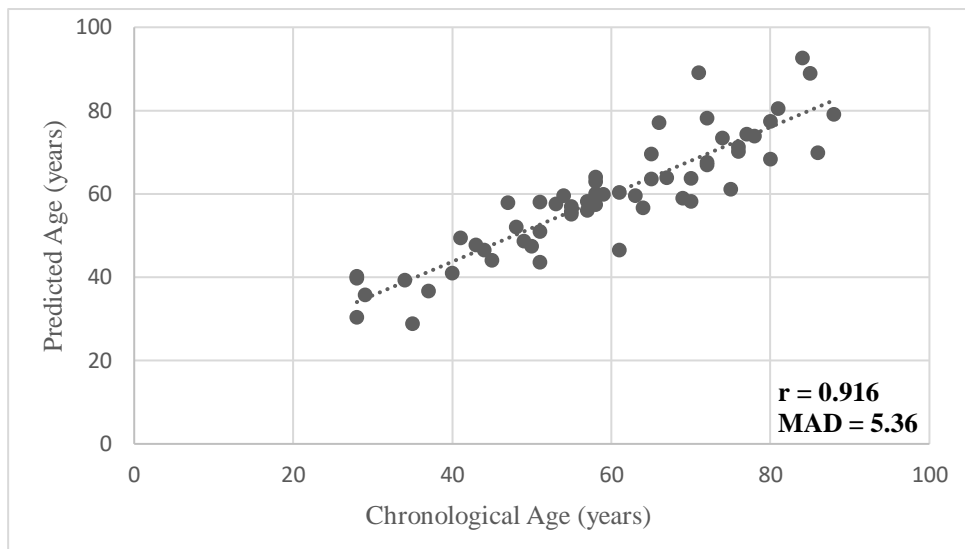


Figure 4.13: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the four CpGs located at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in blood samples from deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Evaluating the MAD values between predicted and chronological ages with the increasing of age using the final APM (with *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes) we observed an increase of MAD values with the increasing of age (**Figure 4.14**).

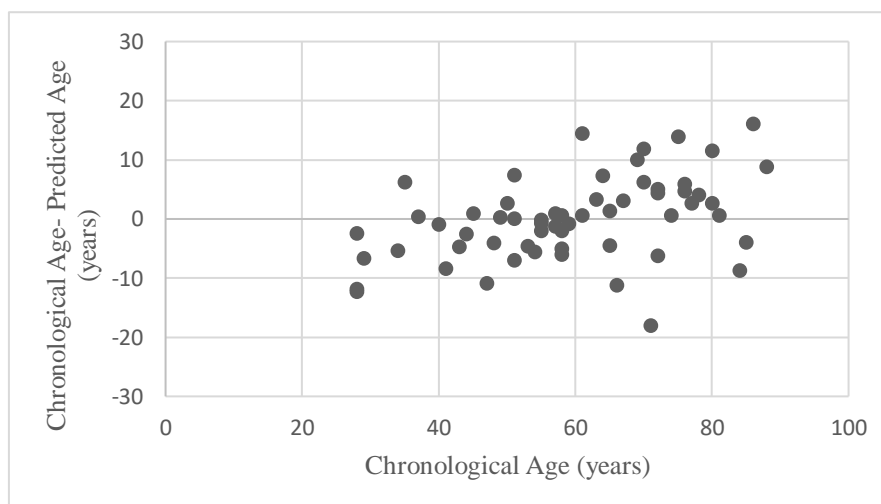


Figure 4.14: Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from deceased individuals.

The largest MAD value was obtained in individuals of the age group 72-88 years old (MAD = 6.24 years), while the smallest MAD was observed in the age group 28-51 years old (MAD = 5.00 years) (Table 4.16; Figure 4.15). Moreover, for younger age categories, we observed higher values of correct predictions (respectively for G1 and G2, 73.7% and 75%) and a lower value in the older age group G3 (68.8%) (Table 4.16).

Table 4.16: MAD between predicted and chronological ages stratified by age group in the training set of 59 blood samples from deceased individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	28-51 years	19	5.00	73.7
G2	52-71 years	24	5.07	75.0
G3	72-88 years	16	6.24	68.8
Overall	28-88 years	59	5.36	72.9

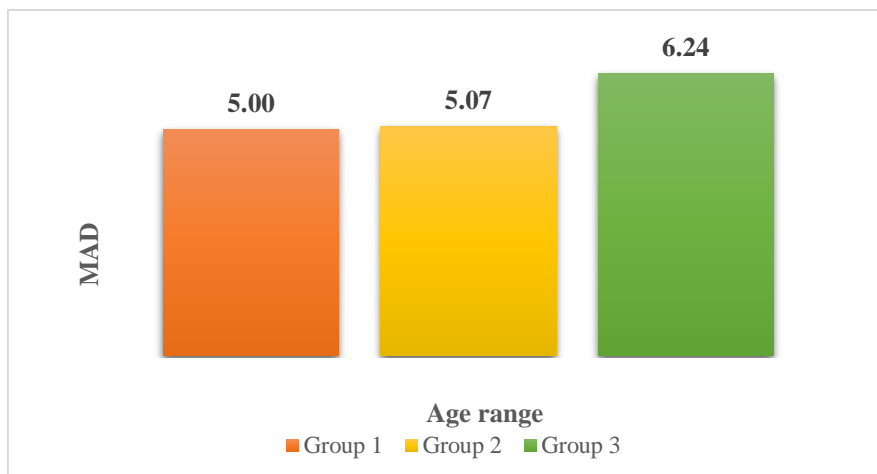


Figure 4.15: MAD from chronological age calculated for each age group in blood samples from deceased individuals. MAD increases with age to age group 3 (72-88 years old). The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in deceased individuals

The 4-fold cross validation allowed to estimate an averaged MAD between predicted and chronological ages obtained for each of the four independent multiple linear

regressions of 6.13 years (RMSE = 6.22), very close to the MAD of 5.36 years (RMSE = 6.94) obtained from the whole data set.

The validation by splitting the sample set into two similar sets of 31 samples and 28 samples (training and validation sets) allowed to obtain an independent MAD value for the training set of 6.15 years (RMSE = 7.35). The model was applied to the validation set and a MAD of 5.66 years was obtained (RMSE = 7.51). Both independent MAD values were very close to the MAD of 5.36 years (RMSE = 6.94) obtained from the whole data set.

3.2.3. Comparison between methodologies

Comparing the obtained data for the sites investigated in both Sanger sequencing and SNaPshot methodologies (*ELOVL2*, *FHL2* and *C1orf132*), very slight differences in age correlation values were observed for *ELOVL2* (R = 0.781 vs. R = 0.791) and *C1orf132* (R = 0.634 vs. R = -0.591) (**Table 4.17**). However, for *FHL2* gene the age correlation value obtained in SNaPshot was higher (R = 0.654) comparing with sequencing (R = 0.431) (**Table 4.17**).

Table 4.17: Comparison of age-correlated values obtained in blood samples from deceased Portuguese individuals through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Blood deceased (24-86 years)		Blood deceased (28-88 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.781	0.610	0.791	0.626
<i>C1orf132</i> Chr1:207823681	-0.634	0.402	-0.591	0.350
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.431	0.186	0.654	0.428

Abbreviations: R, Pearson correlation coefficient.

3.3. Age estimation in blood samples from living and deceased individuals

In the present section, we evaluated the association between chronological age and DNAm levels at *ELOVL2*, *EDARADD*, *FHL2* and *PDE4C* genes in blood samples from both living and deceased individuals that were analyzed using the bisulfite PCR sequencing methodology. The *C1orf132* locus was addressed only for blood samples from deceased individuals and, hence, was excluded from this analysis.

Likewise, we analyzed the association between chronological age and DNAm at the five CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes in blood samples from both living and deceased individuals that were analyzed through the multiplex methylation SNaPshot assay described in Jung *et al.* (2019).

3.3.1. DNAm data obtained in blood samples from living and deceased individuals using bisulfite Sanger sequencing

Correlation between DNAm levels and chronological age

We combined all blood samples from 71 living and 73 deceased individuals evaluated through Sanger sequencing methodology, including samples used in independent validation sets, in a total of 144 blood samples (84 males, 60 females; aged 1-94 years old), for testing the association between DNAm and chronological age. From deceased individuals, three samples did not amplify for *ELOVL2* and *PDE4C* genes and one of these samples for the *EDARADD* gene.

Strong age correlation values were observed for all the CpGs located at *ELOVL2*, for the first three CpGs located at *FHL2* and *PDE4C* genes, and for the CpG5 at *FHL2* and *PDE4C* genes. For the CpG3 located at *EDARADD*, a strong age correlation value was also observed (**Supplementary Table S5**).

Development of an age prediction model (APM)

Testing the association between DNAm levels and chronological age through simple linear regression, the strongest age correlation value within each gene was obtained for *ELOVL2* CpG6 ($R = 0.892$, $P\text{-value} = 7.773 \times 10^{-50}$), explaining 79.5% of the variation in age, *PDE4C* CpG2 ($R = 0.830$, $P\text{-value} = 4.656 \times 10^{-37}$), explaining 68.7% of the variation in age, *FHL2* CpG1 ($R = 0.828$, $P\text{-value} = 1.405 \times 10^{-37}$), explaining 68.4% of the variation in age, and *EDARADD* CpG3 ($R = -0.786$, $P\text{-value} = 3.266 \times 10^{-31}$), explaining 61.5% of the variation in age (**Table 4.18; Supplementary Table S5**). A clear positive correlation between DNAm levels and age was observed for *ELOVL2*

CpG6, *PDE4C* CpG2 and *FHL2* CpG1 markers, and a clear negative correlation was observed for *EDARADD* CpG3 marker (**Supplementary Figure S11**).

Simple APMs using the regression coefficients of these CpG sites allowed to obtain a MAD from chronological age of 8.76 years for *ELOVL2* CpG6, 10.85 years for *PDE4C* CpG2, 11.15 years for *FHL2* CpG1 and 12.68 years for *EDARADD* CpG3 (**Table 4.18; Supplementary Figure S12**).

Table 4.18: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in the overall sample set of blood samples from living and deceased individuals.

Locus	CpG	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	CpG6	Chr6:11044644	141	0.892	0.796	0.795	11.33	7.77×10^{-50}	8.76
<i>PDE4C</i>	CpG2	Chr19:18233133	141	0.830	0.689	0.687	13.99	4.66×10^{-37}	10.85
<i>FHL2</i>	CpG1	Chr2:105399282	144	0.828	0.686	0.684	13.91	1.41×10^{-37}	11.15
<i>EDARADD</i>	CpG3	Chr1:236394382	143	-0.786	0.617	0.615	15.41	3.27×10^{-31}	12.68
<i>Multiple linear regression</i>									
APM (<i>EDARADD</i> CpG3, <i>FHL2</i> CpG1, <i>ELOVL2</i> CpG6 and <i>PDE4C</i> CpG2)			141	0.947	0.897	0.894	8.12	3.47×10^{-66}	6.21

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

The methylation levels of the 141 blood samples from living and deceased individuals for the four most significant age-associated CpG sites (*ELOVL2* CpG6, *PDE4C* CpG2, *FHL2* CpG1 and *EDARADD* CpG3) were put simultaneously in a multiple linear regression analysis to construct the best multi-locus APM. All the predictor variables showed a significant p-value (**Table 4.19**). In concordance, the same pattern of significance was obtained through the stepwise regression analysis using the four selected sites demonstrating that this model should be selected as the final APM. Using the four CpGs, the multiple linear regression statistics showed a very strong age correlation value ($R = 0.947$), highly significant ($P\text{-value} = 3.470 \times 10^{-66}$), explaining 89.4% of the variation in age (**Table 4.18**). The predictive equation developed for estimating age was (**Table 4.19**): $(-50.140) - 74.671 \times \text{DNAm level } EDARADD \text{ CpG3} + 39.409 \times \text{DNAm level } FHL2 \text{ CpG1} + 114.654 \times \text{DNAm level } ELOVL2 \text{ CpG6} + 47.214 \times \text{DNAm level } PDE4C \text{ CpG2}$ (**Figure 4.16**).

Predicted and chronological ages were highly correlated (Spearman correlation coefficient, $r = 0.924$) and the multi-locus APM allowed to obtain a MAD from chronological age of 6.21 years (RMSE = 7.87) (**Figure 4.16; Table 4.18**). Correct predictions were 71.6% assuming that chronological and predicted ages match ± 8 years.

Table 4.19: Statistical parameters obtained in a multiple regression model with the four CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, selected by stepwise regression approach, in the overall set of blood samples from living and deceased individuals.

Marker	Coefficient	P-value
(Intercept)	-50.140	0.001
<i>ELOVL2</i> CpG6	114.654	0.000
<i>FHL2</i> CpG1	39.409	0.001
<i>EDARADD</i> CpG3	-74.671	0.000
<i>PDE4C</i> CpG2	47.214	0.000

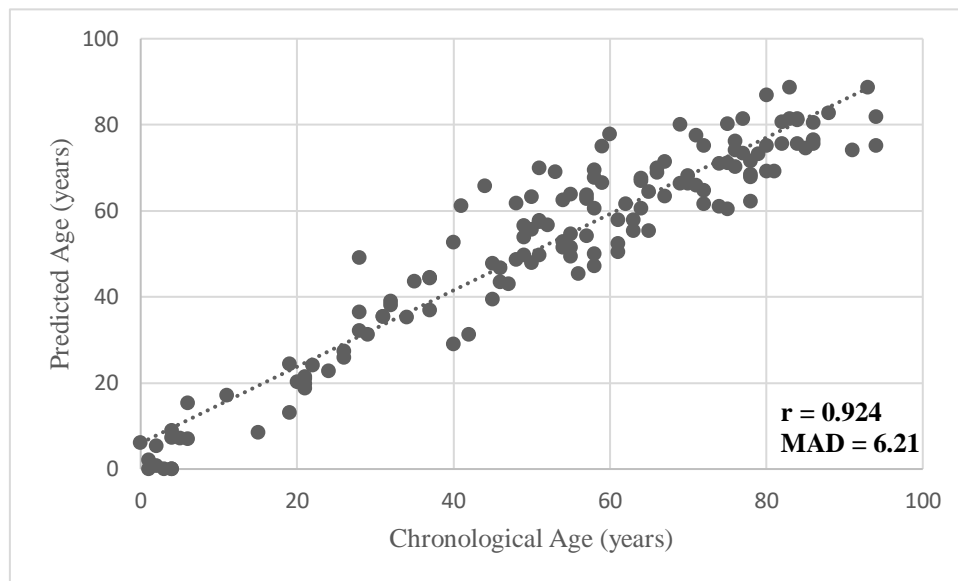


Figure 4.16: Predicted age *versus* chronological age using the four best markers *EDARADD* CpG3, *FHL2* CpG1, *ELOVL2* CpG6 and *PDE4C* CpG2 in blood samples from living and deceased individuals. MAD and Spearman correlation, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Evaluating the differences between ages using the multi-locus APM developed for blood samples from both living and deceased individuals some differences between age categories were observed (**Figure 4.17**).

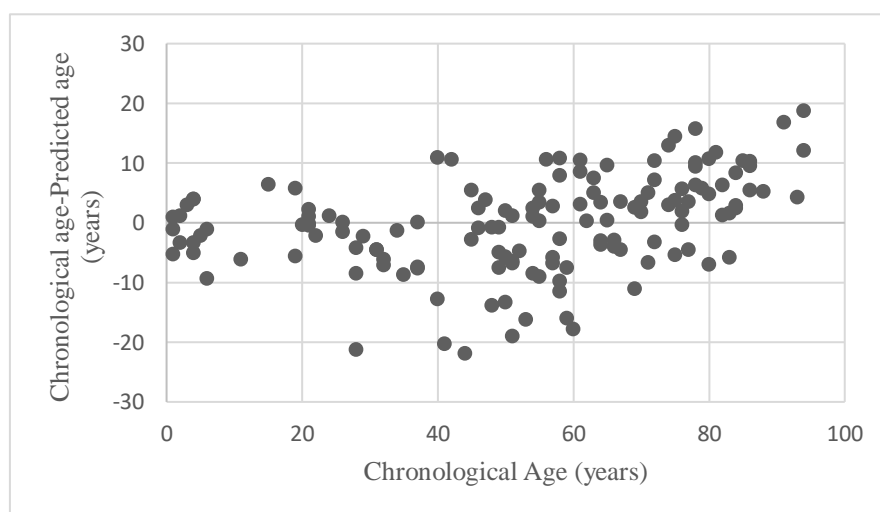


Figure 4.17: Differences between chronological and predicted ages (years) plotted against chronological age (years) in the combined set of blood samples from living and deceased individuals.

To investigate these age-related differences, we split our training set of 141 blood samples in four age groups to estimate MAD and percentage of correct predictions in each age range group (**Table 4.20; Figure 4.18**). Age predictions were considered either correct or incorrect if the predicted age was concordant with the chronological age ± 8 years, according to the standard error of estimate calculated for the final APM (SE = 8.12).

The MAD value is higher in the two older age categories G3 and G4, MAD = 7.79 years and MAD = 6.44 years, respectively. In concordance, the lower percentage of correct predictions was observed in the older age groups G3 and G4 (60.5% and 67.8%, respectively). For younger individuals (G1 and G2), the smaller MAD values and the higher values of correct predictions were observed, MAD = 3.76 years (93.3%) and MAD = 4.33 years (87.5%), respectively (**Table 4.20; Figure 4.18**).

Table 4.20: MAD between predicted and chronological ages stratified by age group in the overall training set of 141 blood samples from living and deceased individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<18 years	15	3.76	93.3
G2	19-39 years	24	4.33	87.5
G3	40-60 years	43	7.79	60.5
G4	>61 years	59	6.44	67.8
Overall	1-94 years	141	6.21	71.6

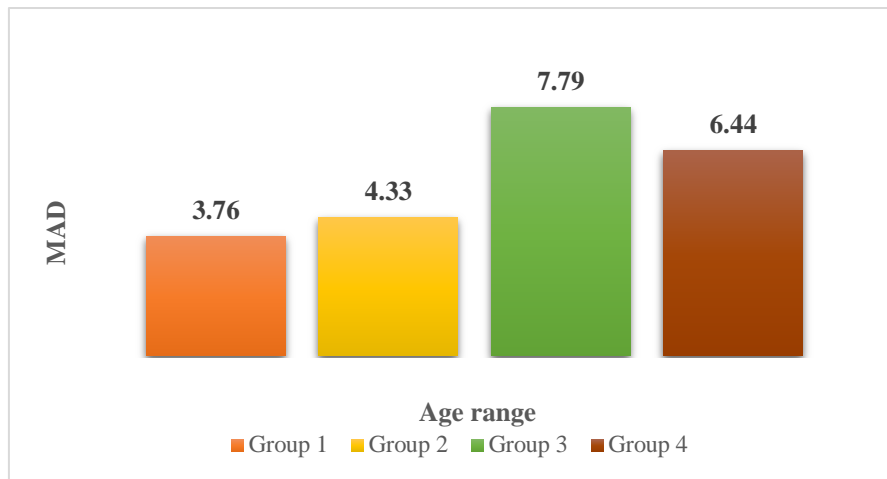


Figure 4.18: MAD from chronological age calculated for each age group in blood samples from living and deceased individuals. The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in living and deceased individuals

A 4-fold cross validation demonstrated a mean MAD value amongst the four validation sets of 6.42 years (RMSE = 6.46), similar to the obtained in the training set, MAD = 6.21 years, showing the accuracy and reproducibility of our developed APM. The validation by splitting the sample into two sets of 71 samples (training set: 39 living; 32 deceased) and 70 samples (validation set: 32 living; 38 deceased) allowed to obtain an independent MAD value for the training set of 5.98 years (RMSE = 7.73). Applying the model on the validation set, a MAD of 6.83 years was obtained (RMSE = 8.35). Both independent MAD values were very close to the overall MAD of 6.21 years.

3.3.2. DNAm data obtained in blood samples from living and deceased individuals using SNaPshot methodology

Correlation between DNAm levels and chronological age

We combined blood samples of 59 living and 62 deceased individuals analyzed through the SNaPshot methodology (in a total of 121 blood samples, 71 males, 50

females; aged 1-94 years old) for assessment of correlation between DNAm levels and age. In blood from living individuals, three samples did not amplify for *ELOVL2* and one sample for *KLF14*; in blood from deceased individuals, one sample did not amplify for *TRIM59*, three samples for *KLF14* and two of these three samples also did not amplify for *C1orf132*.

Among the five genes simultaneously investigated, positive correlations were observed for *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes and a negative correlation was observed for *C1orf132* locus (**Supplementary Figure S13**).

Development of an age prediction model (APM)

The methylation levels of *ELOVL2* locus showed the strongest correlation with chronological age ($R = 0.919$, $P\text{-value} = 7.60 \times 10^{-49}$), explaining 84.4% of the variation in age, followed by *FHL2* ($R = 0.874$, $P\text{-value} = 3.68 \times 10^{-39}$), explaining of 76.2% the variation in age, *C1orf132* ($R = -0.834$, $P\text{-value} = 6.15 \times 10^{-32}$), explaining 69.2% of the variation in age, *TRIM59* ($R = 0.830$, $P\text{-value} = 1.16 \times 10^{-31}$), explaining 68.6% of the variation in age, and *KLF14* ($R = 0.731$, $P\text{-value} = 8.33 \times 10^{-21}$), explaining 53.0% of the variation in age (**Table 4.21; Supplementary Figure S13**). The predicted age of individuals were calculated through the simple linear regression coefficients for the individual markers allowing to obtain MAD values from chronological ages of 7.46 years, 9.23 years, 10.33 years, 10.76 years and 12.76 years for *ELOVL2*, *FHL2*, *C1orf132*, *TRIM59* and *KLF14* genes, respectively (**Supplementary Figure S14**).

Table 4.21: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot in 121 blood samples from living and deceased individuals.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>ELOVL2</i>	Chr6:11044628	118	0.919	0.845	0.844	9.60	7.60×10^{-49}	7.46
<i>FHL2</i>	Chr2:105399282	121	0.874	0.764	0.762	11.81	3.68×10^{-39}	9.23
<i>C1orf132</i>	Chr1:207823681	119	-0.834	0.695	0.692	13.50	6.15×10^{-32}	10.33
<i>TRIM59</i>	Chr3:160450189	120	0.830	0.688	0.686	13.61	1.16×10^{-31}	10.76
<i>KLF14</i>	Chr7:130734355	117	0.731	0.534	0.530	16.43	8.33×10^{-21}	12.76
<i>Multiple linear regression</i>								
APM (<i>ELOVL2</i> , <i>FHL2</i> , <i>C1orf132</i> and <i>TRIM59</i>)		115	0.963	0.928	0.925	6.71	1.03×10^{-61}	4.97

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Combining all the five CpG sites simultaneously in a multiple linear regression analysis to select the best multi-locus APM, a non-significant age correlation value was observed for *KLF14* (P -value = 0.186; **Supplementary Table S6**). In concordance, the stepwise regression analysis also excluded *KLF14* from the model. Therefore, the model for age prediction selecting the four CpG sites at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes used as a final APM, showed the multiple regression age correlation coefficient R of 0.963, explaining 92.5% of the variation in age (corrected $R^2 = 0.925$), highly significant (P -value = 1.025×10^{-61}) (**Table 4.21**). The developed formula obtained with regression coefficients (**Table 4.22**) to calculate age was the following: $17.936 + 66.925 \times \text{DNAm level } ELOVL2 + 52.009 \times \text{DNAm level } FHL2 - 30.886 \times \text{DNAm level } C1orf132 + 44.391 \times \text{DNAm level } TRIM59$. The model showed a strong correlation between predicted and chronological ages (Spearman correlation coefficient, $r = 0.952$), with a MAD of 4.97 years (RMSE = 6.54) (**Figure 4.19**). Correct predictions were 71.3% assuming that chronological and predicted ages match ± 7 years (SE = 6.71).

Table 4.22: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*, selected by stepwise regression approach, in the overall set of blood samples from living and deceased individuals.

Marker	Coefficient	P -value
(Intercept)	17.936	0.001
<i>ELOVL2</i>	66.925	0.000
<i>FHL2</i>	52.009	0.000
<i>C1orf132</i>	-30.886	0.000
<i>TRIM59</i>	44.391	0.000

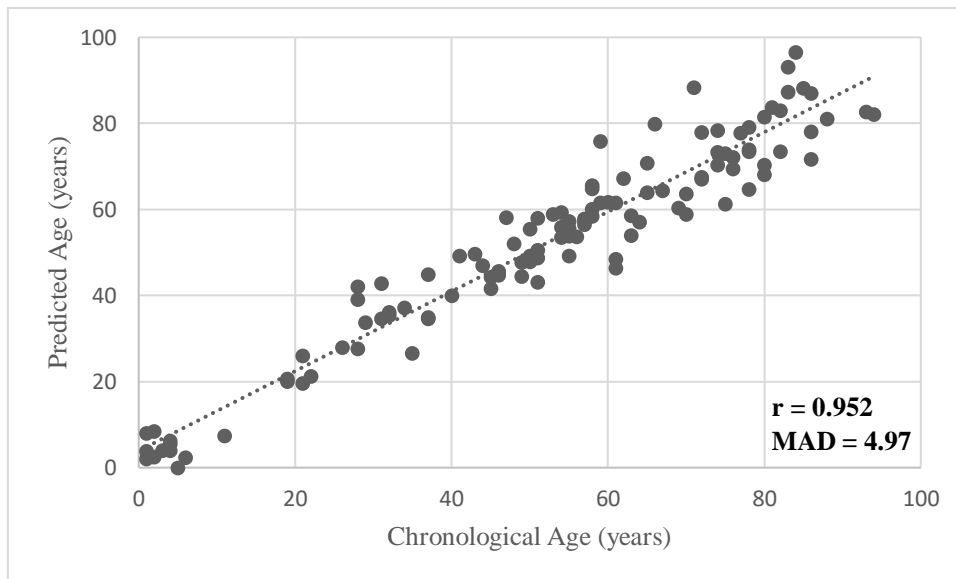


Figure 4.19: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the four CpGs located at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in blood samples from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Considering the final APM with *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes developed for blood from living and deceased, we noted some differences in DNAm with the increasing of age (**Figure 4.20**).

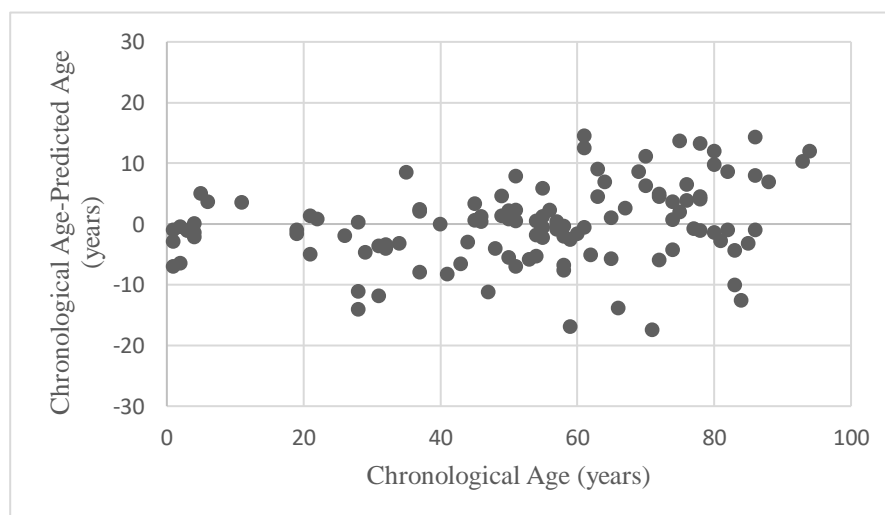


Figure 4.20: Differences between chronological and predicted ages (years) plotted against chronological age (years) in the combined set of blood samples from living and deceased individuals.

Testing the differences in four age categories, the lower MAD values were observed for the first three age categories, G1, G2 and G3 (Table 4.23; Figure 4.21). The older individuals (G4, >61 years old) showed the higher MAD value (MAD = 6.78 years) and the lower percentage of correct predictions (56.5%) (Table 4.23; Figure 4.21).

Table 4.23: MAD between predicted and chronological ages stratified by age group in the overall training set of 115 blood samples from living and deceased individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<18 years	12	2.88	91.7
G2	19-39 years	19	4.68	73.7
G3	40-60 years	38	3.58	81.6
G4	>61 years	46	6.78	56.5
Overall	1-94 years	115	4.97	71.3

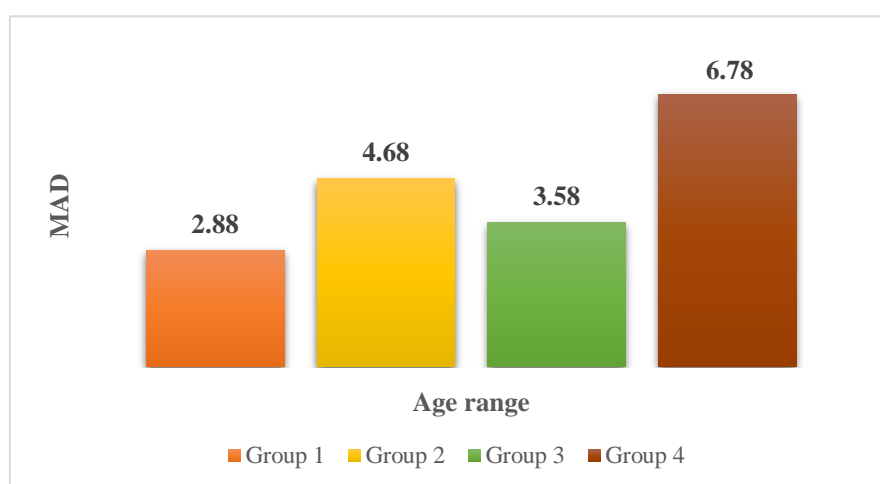


Figure 4.21: MAD from chronological age calculated for each age group in blood samples from living and deceased individuals. The MAD is printed on top of each respective age range.

Validation of the multi-locus APM developed in living and deceased individuals

A 4-fold cross validation using the whole dataset of 115 individuals allowed to estimate an averaged MAD of 5.25 years (RMSE = 5.29), close to the MAD of 4.97 years. The validation by splitting the sample into two sets of 58 samples (training set: 27 living; 31 deceased) and 57 samples (validation set: 29 living; 28 deceased) allowed to obtain an

independent MAD value for the training set of 5.35 years (RMSE = 6.72). Applying the model on the validation set, a MAD of 5.35 years was obtained (RMSE = 7.18). Both independent MAD values were very close to the MAD of 4.97 years for the overall training set.

3.3.3. Comparison between methodologies

Comparing our obtained data for Portuguese people using Sanger sequencing and SNaPshot methodologies, for the sites investigated in both methodologies (*ELOVL2*, Chr6:11044628 and *FHL2*, Chr2:105399282) in the overall sample set of blood samples from living and deceased individuals, the age correlation values obtained for these two sites revealed strong or very strong age correlation values, showing very slight differences: for *ELOVL2* (R = 0.872 vs. R = 0.919) and for *FHL2* (R = 0.828 vs. R = 0.874) (Table 4.24).

Table 4.24: Comparison of age-correlated values obtained in blood samples from living and deceased Portuguese individuals through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Blood living and deceased (1-94 years)		Blood living and deceased (1-94 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.872	0.760	0.919	0.845
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.828	0.686	0.874	0.764

Abbreviations: R, Pearson correlation coefficient.

3.4. Applicability of the developed APMs for blood samples from living and deceased individuals

To further analyse the applicability of the constructed APMs, we tested the obtained models in blood samples from living individuals through both methodologies of Sanger sequencing and SNaPshot in deceased individuals and *vice versa*.

3.4.1. Applying the developed APMs in blood samples from living individuals to deceased individuals

The final multi-locus equation developed using blood samples from living Portuguese individuals (**Table 4.3**) applied to methylation information captured from blood samples from deceased individuals through Sanger sequencing methodology allowed to obtain a MAD from chronological age of 9.72 years (RMSE = 12.25), revealing a lower accuracy when compared with the MAD obtained in living individuals (MAD = 5.35 years, **Table 4.2**). Spearman correlation between predicted and chronological age was 0.792 (**Figure 4.22**).

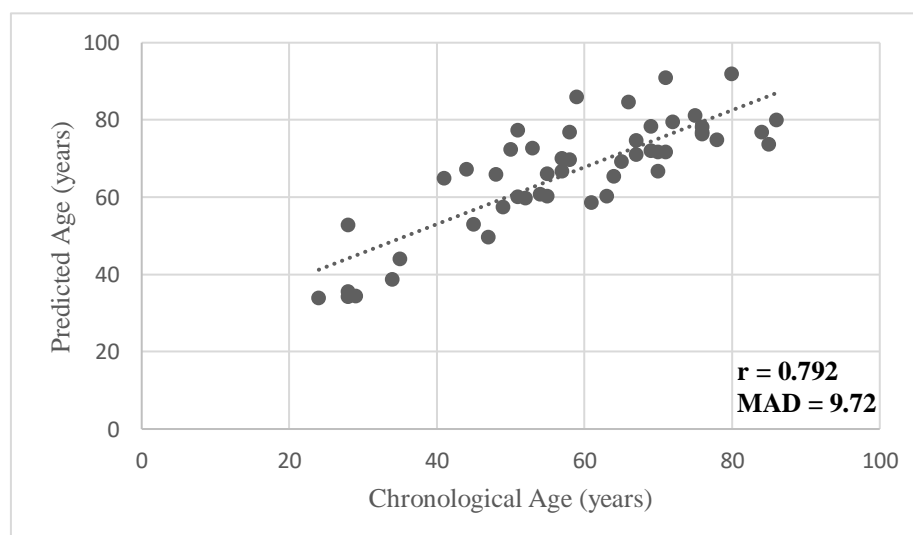


Figure 4.22: Predicted age *versus* chronological age in a test sample set of blood samples from deceased individuals evaluated through Sanger sequencing methodology. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Applying the equation developed for blood samples from living individuals (**Table 4.8**) through SNaPshot to methylation information captured from blood samples of deceased individuals, the obtained MAD value was 7.84 years (RMSE = 9.96), revealing also lower accuracy when compared with the MAD obtained in living individuals (MAD = 4.25 years; **Table 4.7**). Spearman correlation between predicted and chronological age was 0.862 (**Figure 4.23**).

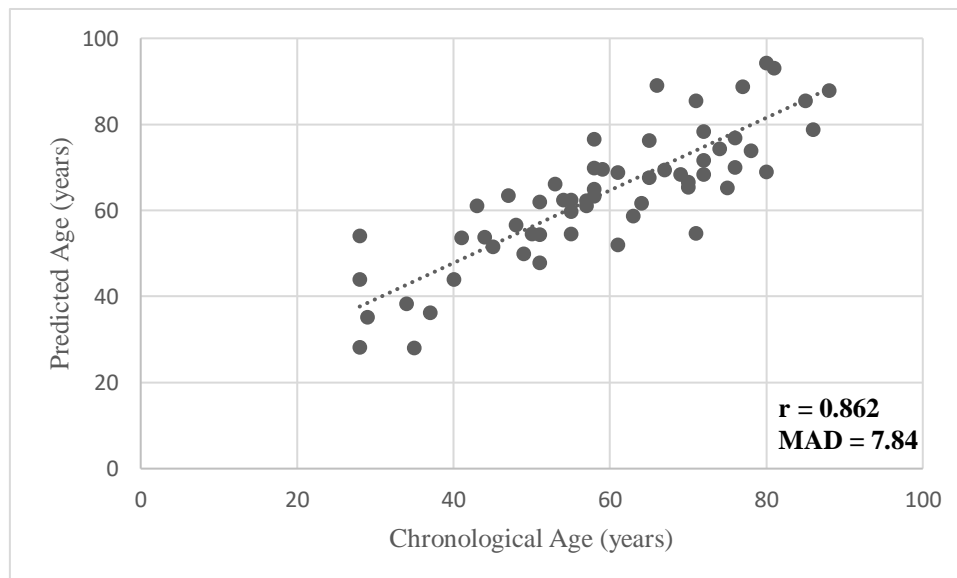


Figure 4.23: Predicted age *versus* chronological age in a test sample set of blood samples from deceased individuals evaluated through SNaPshot methodology. MAD and Spearman correlation coefficient, r , are plotted on the chart.

3.4.2. Applying the developed APMs in blood samples from deceased individuals to living individuals

As CpG sites in *C1orf132* gene were not investigated in blood from living individuals through bisulfite sequencing, we cannot test the reproducibility of the previous developed multi-locus APM using blood from deceased individuals with *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *C1orf132* CpG1 and *PDE4C* CpG2. Hence, we constructed a new APM for blood samples from deceased individuals

considering the best CpG sites located at *ELOVL2* (CpG4), *PDE4C* (CpG2), *EDARADD* (CpG3) and *FHL2* (CpG2). The new multiple linear multi-locus regression model developed for the 49 blood samples from deceased individuals, showed a strongly age correlation value ($R = 0.871$), highly significant ($P\text{-value} = 5.01 \times 10^{-13}$), explaining 73.6% of the variation in age (**Table 4.25**). The new predictive equation based on regression coefficients (**Supplementary Table S7**) was: $(-40.063) - 74.494 \times \text{DNAm level } EDARADD \text{ CpG3} + 59.531 \times \text{DNAm level } FHL2 \text{ CpG2} + 100.423 \times \text{DNAm level } ELOVL2 \text{ CpG4} + 26.479 \times \text{DNAm level } PDE4C \text{ CpG2}$. Predicted and chronological ages of the 49 deceased individuals were highly correlated (Spearman correlation coefficient, $r = 0.849$) with a MAD from chronological age of 6.42 years (RMSE = 7.94) (**Figure 4.24**).

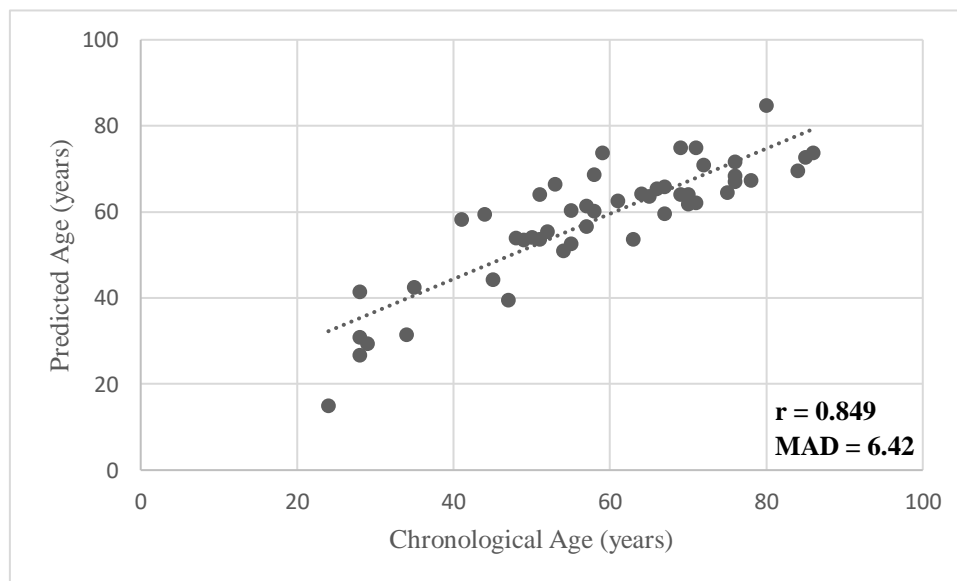


Figure 4.24: Predicted age *versus* chronological age using the APM developed with the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3 and *PDE4C* CpG2 genes in the training set of deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Table 4.25: Multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test the association between the DNAm levels and chronological age in blood samples from deceased individuals using the Sanger sequencing methodology.

<i>Multiple linear regression</i>	N	R	R²	Corrected R²	SE	P-value	MAD
APM (<i>ELOVL2</i> CpG4, <i>FHL2</i> CpG2, <i>EDARADD</i> CpG3 and <i>PDE4C</i> CpG2)	49	0.871	0.758	0.736	8.38	5.01×10^{-13}	6.42

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Testing the reproducibility of the new APM with *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3 and *PDE4C* CpG2 in blood samples from living individuals a MAD of 6.10 years was obtained (**Figure 4.25**), revealing similar accuracy when comparing with the APM developed in deceased individuals (MAD = 6.42 years; **Table 4.25**). Spearman correlation value between predicted and chronological ages was 0.974.

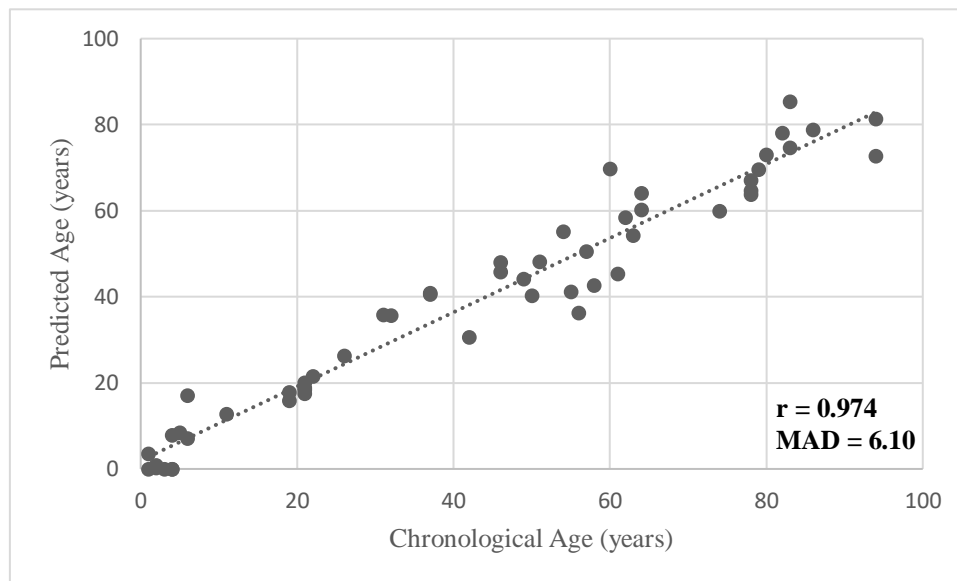


Figure 4.25: Predicted age *versus* chronological age in a test sample set of 53 blood samples from living individuals using the new model developed using blood from deceased individuals with the markers *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3 and *PDE4C* CpG2 markers. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Testing the reproducibility of our final APM developed with methylation information of blood samples from deceased individuals through SNaPshot methodology (**Table 4.15**) in the set of healthy Portuguese individuals allowed to obtain a MAD from chronological age of 5.40 years (RMSE = 6.88), which is similar to the accuracy of the model developed in deceased (MAD = 5.36 years; **Table 4.14**). Spearman correlation between predicted and chronological ages was 0.967 (**Figure 4.26**).

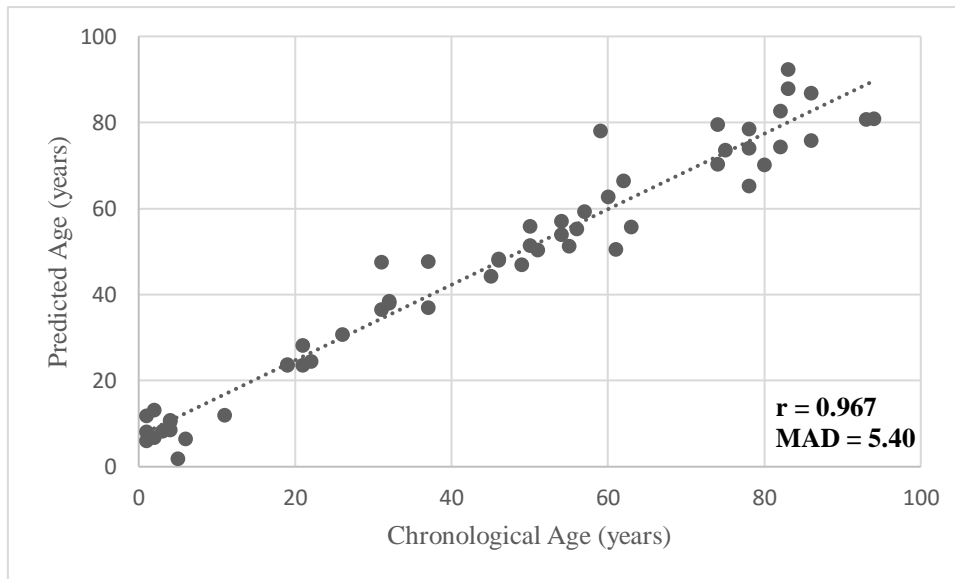


Figure 4.26: Predicted age *versus* chronological age in a test sample set of blood samples from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

4. Discussion

Age estimation plays a relevant role in forensic science since it can be very useful in criminal, legal and civil investigations, including for the establishment of criminal responsibility or in immigration cases (Parson *et al.*, 2018; Nuzzolese and Di Vella, 2019). Moreover, prediction of age from biological evidences can be very useful for identification purposes of human remains from mass disasters or to solve crimes by limiting the search range of unknown suspects (Jung *et al.*, 2017; Parson *et al.*, 2018).

In recent years, DNAm age has arisen as a promising tool for forensic age estimations. Previous forensic studies, most of them with blood samples of living individuals, using different loci and different number of markers, covering different age ranges and using different methodologies, gave high accurate APMs with values of MAD from chronological age between 3.5 to 7.5 years (Weidner *et al.*, 2014; Bekaert *et al.*, 2015b; Zbieć-Piekarska *et al.*, 2015a; Cho *et al.*, 2017; Naue *et al.*, 2017; Aliferi *et al.*,

2018; Jung *et al.*, 2019). To the best of our knowledge, only four studies have focused in blood samples from deceased individuals for development of APMs based on DNAm among a number of different loci including *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Naue *et al.*, 2018; Pfeifer *et al.*, 2020).

In this study, we used the bisulfite PCR sequencing method to analyze the methylation patterns of CpG markers in four well-known age-associated loci (*ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*) in a total number of 144 blood samples from living (training set of 53 individuals; validation set of 18 individuals) and deceased (training set of 51 individuals; validation set of 22 individuals). Additionally, DNAm of the *C1orf132* gene was also analyzed by using the same methodology in blood samples from deceased individuals. The selection of these markers was made based on their high age-correlated values observed in several previous studies in blood and other tissue types (Garagnani *et al.*, 2012; Weidner *et al.*, 2014; Florath *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Xu *et al.*, 2015; Hamano *et al.*, 2016; Giuliani *et al.*, 2016; Freire-Aradas *et al.*, 2016; Kananen *et al.*, 2016; Park *et al.*, 2016; Spólnicka *et al.*, 2017; Fleckhaus *et al.*, 2017; Cho *et al.*, 2017; Thong *et al.*, 2017; Naue *et al.*, 2017, 2018; Freire-Aradas *et al.*, 2017; Jung *et al.*, 2019; Pfeifer *et al.*, 2020).

Moreover, for better application of developed approaches to forensic casework sample analysis it is important to further test the proposed markers and methodologies to establish consistency between populations and laboratories (Cho *et al.*, 2017; Daunay *et al.*, 2019). Hence, we tested the experiments proposed by Jung *et al.* (2019) using the SNaPshot methodology and the five CpG sites located in *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes in a set of 121 samples from Portuguese individuals (59 blood samples of healthy individuals and 62 blood samples from deceased individuals).

In the present study, a total of 37 CpGs (*ELOVL2*: 9 CpGs; *FHL2*: 12 CpGs; *PDE4C*: 12 CpGs; and *EDARADD*: 4 CpGs) was evaluated through Sanger sequencing in a training set of 53 blood samples from healthy Portuguese individuals. The same methodology was used to evaluate a total of 43 CpGs (*ELOVL2*: 9 CpGs; *FHL2*: 12 CpGs; *PDE4C*: 12 CpGs; *EDARADD*: 4 CpGs; and *C1orf132*: 6 CpGs) in a training set of 51 blood samples from deceased individuals. The training set of blood samples from living individuals revealed the highest age correlation value per locus for *FHL2* CpG3 ($R = 0.940$), *ELOVL2* CpG6 ($R = 0.936$), *EDARADD* CpG3 ($R = -0.888$) and *PDE4C* CpG2 ($R = 0.852$). Blood samples from deceased individuals revealed the highest age correlation value per locus for *ELOVL2* CpG4 ($R = 0.785$), *C1orf132* CpG1 ($R = -0.634$), *EDARADD* CpG3 ($R = -0.621$), *PDE4C* CpG2 ($R = 0.592$) and *FHL2* CpG2 ($R = 0.465$). In the overall training set of blood samples from both living and deceased individuals, the highest age correlation value per locus was observed for *ELOVL2* CpG6 ($R = 0.892$), *PDE4C* CpG2 ($R = 0.830$), *FHL2* CpG1 ($R = 0.828$) and *EDARADD* CpG3 ($R = -0.786$).

Evaluating simultaneously through multiple linear regression the methylation information of the highest age-correlated CpG sites observed in blood samples from living individuals, a final four-locus APM was constructed, revealing good accuracy with a MAD from chronological age of 5.35 years, explaining 94.1% of the variation in age. In the training set of blood samples from deceased individuals, the final model built with the five highest age-associated CpG sites revealed the best results in age correlation, explaining 76.3% of the variation in age, and a MAD of 6.08 years. Moreover, the multi-locus APM constructed with the methylation information of the four best age-associated CpG sites observed in the overall training set of blood samples (from both living and deceased individuals), allowed to obtain a MAD from the chronological age of 6.21 years, explaining 89.4% of the variation in age. This is a similar value in comparison with the

obtained for both independent models (MAD = 5.35 years in living with four CpGs, and MAD = 6.08 years in deceased with five CpGs).

The MAD and age correlation values reported in our study using blood samples from living individuals are similar to the obtained in other APMs using pyrosequencing which has been the preferred method by forensic epigeneticists to assess site specific DNAm levels in age predictions, but also other methods (Garagnani *et al.*, 2012; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Thong *et al.*, 2017; Cho *et al.*, 2017). Garagnani *et al.* (2012), using Sequenom's EpiTYPER assay, reported high age correlation values (Pearson correlation) for CpGs located at *ELOVL2* Chr6:11044644 (cg16867657), Chr6:11044661 (cg21572722) and Chr6:11044655 (cg24724428) (0.91, 0.89 and 0.85, respectively), which are similar to the values obtained in our study for the same sites: *ELOVL2* CpG6, *ELOVL2* CpG9 and *ELOVL2* CpG8 (0.936, 0.860 and 0.850, respectively). For CpGs located at *FHL2*, also a similar value was obtained in Garagnani *et al.* (2012) for Chr2:105399310 (cg22454769) and in our study for the same position, *FHL2* CpG6 (0.92 and 0.848, respectively). Bekaert *et al.* (2015a) using the pyrosequencing methodology reported a MAD of 3.75 years using a multivariate quadratic model with CpGs located at *ELOVL2*, *EDARADD*, *PDE4C* and *ASPA* genes. Interestingly, the selected CpG located at *EDARADD* demonstrated a $R^2 = 0.62$, similar to the value observed in our study for the same position *EDARADD* CpG3 ($R^2 = 0.788$). Zbieć-Piekarska *et al.* (2015b) evaluated DNAm patterns of *ELOVL2*, *Clorf132*, *TRIM59*, *KLF14* and *FHL2* loci by pyrosequencing revealing a high age correlation value ($R = 0.971$) and a MAD of 3.4 years. This group reported strong age correlation values (Spearman correlation value) in CpGs located at *ELOVL2* Chr6:11044661 and Chr6:11044644 (0.828 and 0.847, respectively) and CpGs located at *FHL2* Chr2:105399288 and Chr2:105399291 (0.843 and 0.826, respectively). In concordance,

our study revealed also similar strong age correlation values for the same sites in *ELOVL2* (0.849 and 0.926, respectively) and in *FHL2* (0.954 and 0.951, respectively). Thong *et al.* (2017) and Cho *et al.* (2017) using pyrosequencing, reported strong age correlation values (Spearman correlation value) for the CpG located at *FHL2* (Chr2:105399291) (0.905 and 0.878 respectively) and for the CpG located at *ELOVL2* (Chr6:11044644) (0.951 and 0.874, respectively); these values were similar to those obtained in our study for the same positions *FHL2* CpG3 (0.951) and *ELOVL2* CpG6 (0.926). Moreover, both studies reported high accuracy with a MAD around of 3.3 years.

In the present study we also tested the experiments proposed by Jung *et al.* (2019) using a SNaPshot assay in an independent set of 59 blood samples of healthy Portuguese individuals using the same markers and methodology. Moreover, we evaluated DNAm levels through the SNaPshot method in a set of 62 blood samples from deceased individuals. Strong age correlation values were obtained for the five CpG sites at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes in living individuals ($0.791 \leq R \leq 0.951$). For deceased individuals, lower age-correlated values, although significant, were obtained ($0.568 \leq R \leq 0.791$). For both the training sets, the DNAm of *ELOVL2* showed the strongest age correlation value. The proposed final APM in living Portuguese individuals was constructed with the three CpGs at *ELOVL2*, *FHL2* and *C1orf132* genes that showed age correlation significance in the multiple regression analysis. This APM exhibited an age prediction accuracy (MAD = 4.25 years) similar to the model of Jung *et al.* (2019) constructed with all the five CpGs in genes *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14* and *FHL2* in the Korean population demonstrating high age prediction accuracy (MAD = 3.174). Thus, our study replicated the age association test for the SNaPshot assay of Jung *et al.* (2019) in an independent validation set of living Portuguese individuals,

evidencing that the performance of these age-related markers through the SNaPshot methodology is relatively consistent across different populations.

When we compare DNAm levels obtained from the multiplex methylation SNaPshot assay in blood samples from the Portuguese and the Koreans (Jung *et al.*, 2019) all the CpG sites showed strong or very strong correlation with age ($0.763 \leq R \leq 0.951$), except for *C1orf132* ($R = -0.637$) in Koreans. However, some differences can be found in the extent of the age association for the targeted loci comparing our data with that obtained in the Korean population. The higher correlation with age was obtained for the CpG sites at *FHL2* in Koreans ($R = 0.893$) and *ELOVL2* in Portuguese ($R = 0.951$). On the other hand, the lowest correlation value was found for *KLF14* in Portuguese, and for *C1orf132* in Koreans. Moreover, in the multivariate analysis the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values ($P\text{-value} > 0.05$) in Portuguese individuals, while in Koreans all the five CpG sites showed significant age correlation values ($P\text{-value} < 0.001$) (Jung *et al.*, 2019). Consequently, the APM proposed by Jung *et al.* (2019), included DNAm at the five genes, while our model included only three genes. Moreover, when we tested the original model of Jung *et al.* (2019) with methylation data from living Portuguese individuals, a higher MAD value of 15.26 years was obtained, revealing difference in applicability of the model of Jung *et al.* (2019) (developed for Koreans) to the set of Portuguese individuals, possible due to differences in ancestry. Similarly, the study of Cho *et al.* (2017) in the Korean population, replicating a study of Zbieć-Piekarska *et al.* (2015b) in the Polish population using pyrosequencing, found age correlation differences in specific markers. Such kind of differences suggests the possibility that some effects in methylation levels may be population specific. This phenomenon supports the notion that specific markers can be more adequate to different population groups to explain age-related DNAm variance, which points to the usefulness

of replication and validation studies of proposed markers and genotyping methods in different populations and datasets before forensic applications.

Regarding blood from deceased individuals, the final multi-locus APM constructed with the predictor variables *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* revealed a strong age correlation ($R = 0.899$) and exhibiting an accurate age prediction with a MAD from chronological age of 5.36 years. For the whole sample set, addressing both living and deceased individuals, the final multi-locus model for age prediction constructed with four CpGs located at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes obtained from the stepwise regression analysis, showed a high age prediction accuracy (MAD = 4.97 years), explaining 92.5% of the variation in age. This MAD value is comparable to those obtained in blood samples from deceased (MAD = 5.36 years) and living individuals (MAD = 4.25 years).

In our study, CpGs from the *ELOVL2* gene were always the strong age-correlated in comparison with all the other investigated markers (*FHL2*, *EDARADD*, *PDE4C*, *C1orf132*, *TRIM59* and *KLF14*), either using SNaPshot or Sanger sequencing. Using the Sanger sequencing methodology, in blood samples from living individuals all the CpGs located at *ELOVL2* revealed strong or very strong correlation values with chronological age ($R \geq 0.850$). In blood samples from deceased individuals only the *ELOVL2* gene showed high significant age-correlation values for all the selected CpG sites reflecting a similar strength on the change in DNAm with chronological age ($R \geq 0.663$). Similarly, for the SNaPshot methodology, the CpG site in the *ELOVL2* showed the strongest age correlation in living ($R = 0.951$) and deceased ($R = 0.791$). This result is in concordance with previous reports showing that *ELOVL2* gene is a stable age-associated marker that exhibits the most strong age-related changes, being currently used in several APMs in different tissues such as whole blood, teeth, saliva, buccal swabs and bone samples

(Garagnani *et al.*, 2012; Hannum *et al.*, 2013; Weidner *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Xu *et al.*, 2015; Giuliani *et al.*, 2016; Spólnicka *et al.*, 2017; Naue *et al.*, 2017, 2018; Sliker *et al.*, 2018; Jung *et al.*, 2019; Gopalan *et al.*, 2019; Márquez-Ruiz *et al.*, 2020; Pfeifer *et al.*, 2020). In a recent work, Bacalini *et al.* (2017) demonstrated that most tissues show *ELOVL2* hypermethylation with age. The authors demonstrated that this hypermethylation is associated with in vitro cell replication rather than with senescence, indicating that *ELOVL2* methylation is a marker of cell divisions occurring during human aging. Moreover, a study by Spólnicka *et al.* (2017) emphasizes the high utility of the *ELOVL2* marker for prediction of chronological age in forensics by showing unchanged prediction accuracy in individuals affected by three diseases (Graves's disease and early or late onset Alzheimer's disease).

In our study, bisulfite sequencing and SNaPshot methodologies showed that the prediction accuracy depends on the chronological age of samples, both in living and deceased individuals. We obtained the largest MAD values in older ages, which was concordant with previous studies (Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015a, 2015b; Hamano *et al.*, 2016; Pfeifer *et al.*, 2020). This means that individual differences in the rate of methylation change occur with age, being slight in youths and accumulating with age, enabling that the APMs are more accurate in younger than in older individuals (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Pfeifer *et al.*, 2020). The fact that DNAm patterns predict age with more accuracy in younger than in older individuals suggests increased inter-individual variation within older people, possibly due to environmental, diseases and stochastic factors (Jaenisch and Bird, 2003; Boks *et al.*, 2009; Heyn *et al.*, 2012; Spólnicka *et al.*, 2017). In spite of this, DNAm age is accepted as an “epigenetic clock”, reflecting stable DNAm changes across the genome, nevertheless capturing

aspects of the biological age of the individual (Hannum *et al.*, 2013; Horvath, 2013; Zbieć-Piekarska *et al.*, 2015b; Horvath and Raj, 2018; Horvath *et al.*, 2018).

Another relevant issue is the possible influence of sex in DNAm levels of age-correlated markers. To date, there is no consensus for a relationship between age-associated DNAm levels and sex (Hannum *et al.*, 2013; Bekaert *et al.*, 2015a, 2015b; Huang *et al.*, 2015; Zbieć-Piekarska *et al.*, 2015b; Freire-Aradas *et al.*, 2018). Our study showed for blood samples of living individuals very slight sex differences or no statistical differences, in accordance with previous studies (Bekaert *et al.*, 2015a; Huang *et al.*, 2015; Zbieć-Piekarska *et al.*, 2015b; Freire-Aradas *et al.*, 2018; Daunay *et al.*, 2019). Therefore, it seems that sex did not influence DNAm levels of the age-related CpG sites, at least for those selected in our study. No statistical comparison between males and females was made for deceased individuals in our study, because only 15 females were evaluated by Sanger sequencing and 13 females through the SNaPshot assay.

For forensic casework application of developed APMs the evaluation of a putative influence of *postmortem* changes in DNAm levels could be an important topic. However, to the best of our knowledge, only four studies have focused in blood samples from deceased individuals for development of APMs based on DNAm among a number of different loci including *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, *KLF14* and *TRIM59* genes (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Naue *et al.*, 2018; Pfeifer *et al.*, 2020). When comparing blood samples from living and deceased individuals, Bekaert *et al.* (2015a), Hamano *et al.* (2016) and Pfeifer *et al.* (2020) suggested similar distributions of DNAm patterns. Bekaert *et al.* (2015a) using the pyrosequencing methodology, analyzed 169 blood samples of deceased individuals (including 37 blood samples from living individuals) (age range 0–91 years) for *ELOVL2*, *PDE4C*, *EDARADD* and *ASPA* genes. The authors proposed an overall prediction model using the four best age-associated CpG

sites from each gene, which demonstrated to be highly accurate, explaining 95% of the variation in age with a MAD of 3.75 years. Interestingly, the highest age-correlated CpG sites from *ELOVL2* (CpG4; Chr6:11044640) and *EDARADD* (CpG3; Chr1:236394382) found in training set of the deceased Portuguese individuals captured by Sanger sequencing methodology correspond to the same best positions observed by Bekaert *et al.* (2015a). The work of Hamano *et al.* (2016), using the MS-HRM method to address the methylation levels of *ELOVL2* and *FHL2* genes, revealed similar distributions of DNAm levels in 22 living blood samples and 52 dead blood samples. However, the authors suggested that potential differences in methylation patterns between living and dead samples could be ignorable by the MS-HRM method. Naue *et al.* (2018) investigated through massively parallel sequencing 13 previously selected age-dependent loci in tissues such as brain, bone, muscle, buccal swabs and whole blood of 29 deceased individuals (aged 0–87 years old), including amongst others the *ELOVL2* locus. All the analyzed markers in blood (*DDO*, *ELOVL2*, *F5*, *GRM2*, *HOXC4*, *KLF14*, *LDB2*, *NKIRAS2*, *RPA2*, *SAMD10*, *TRIM59*, *MEIS1* and *ZYG11A*) showed a comparable age-dependency in comparison to a previous study using whole blood of living individuals (Naue *et al.*, 2017). Interestingly, the *ELOVL2* position (Chr6:11044644; cg16867657) showed a Pearson correlation coefficient of 0.88, which is similar to the value ($R = 0.76$) obtained in our training set of deceased individuals using Sanger sequencing for the same position (CpG6). In concordance, Pfeifer *et al.* (2020) demonstrated no significant differences in DNAm levels of *PDE4C*, *ASPA*, *EDARADD* and *ELOVL2* genes between blood samples from living and deceased individuals.

In regards to our study, potential differences in methylation status between samples from living and deceased individuals were observed since the highest age-correlated CpGs were different in some genes between both groups using Sanger

sequencing methodology: in living individuals the best age-associated markers were *ELOVL2* CpG6, *EDARADD* CpG3, *PDE4C* CpG2 and *FHL2* CpG3, whereas in deceased individuals the strong age-associated markers were *ELOVL2* CpG4, *EDARADD* CpG3, *PDE4C* CpG2 and *FHL2* CpG2. Moreover, the correlation between DNAm and age obtained in genes *ELOVL2*, *EDARADD*, *FHL2* and *PDE4C* is lower in blood samples of deceased individuals ($0.30 \leq R \leq 0.79$) vs. living individuals ($0.41 \leq R \leq 0.94$), with *ELOVL2* revealing the higher and most similar DNAm levels in both groups. Through the SNaPshot methodology, the individual analysis of DNAm levels also showed higher age correlation values in living individuals ($0.79 \leq R \leq 0.95$) in relation to blood samples from deceased ($0.57 \leq R \leq 0.79$). Moreover, in the multivariate analysis, the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values in living individuals, while in deceased individuals only the CpG at *KLF14* showed a non-significant p-value. All these data could be explained by *postmortem* changes that can alter the methylation status among specific loci. After death, there is a decay of biological functions and the organism suffers some alterations. Consequently, it is possible that some of these changes influence DNAm levels. These DNAm changes after death suggest that a specific APM developed for blood samples from living individuals cannot accurately be applied for blood samples from deceased individuals. In concordance, when we applied the final APM built for living individuals through Sanger sequencing (APM with *ELOVL2* CpG6, *EDARADD* CpG3, *FHL2* CpG3 and *PDE4C* CpG2; MAD = 5.35 years) to the independent set of blood samples from deceased individuals, we obtained a higher value of MAD (9.72 years), which represents a decrease of the model accuracy in this set of samples. In particular, for the selected sites included in APM for living individuals, strong or very strong age correlation values were obtained ($0.852 \leq R \leq 0.940$), whereas in deceased individuals we observed for the same sites, lower age correlation values ($0.459 \leq$

$R \leq 0.764$). In the same way, when we applied the final multi-locus APM developed for living individuals using the SNaPshot methodology (with *ELOVL2*, *FHL2* and *C1orf132*) to the methylation information captured in the set of deceased individuals using the same methodology and the same markers, a MAD of 7.84 years was observed. This value is higher when comparing with the accuracy obtained in the model developed for living individuals (MAD = 4.25 years).

For testing APMs developed in blood samples from deceased individuals to age prediction in living individuals, we developed an additional APM for blood samples from deceased individuals using Sanger sequencing with methylation information from *ELOVL2* CpG4, *EDARADD* CpG3, *PDE4C* CpG2 and *FHL2* CpG2 (removing the *C1orf132* marker that was not analyzed in living individuals), revealing a MAD from chronological age of 6.42 years. When this model was tested in the training set of living individuals a similar high accuracy was observed (MAD = 6.10 years). In concordance, when we applied the model developed for blood samples from deceased individuals (with *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59*) using SNaPshot methodology (MAD = 5.36 years) to the set of living individuals, a similar MAD of 5.40 years was obtained. Thus, we hypothesized that APMs developed using methylation information captured in blood samples from deceased individuals (using Sanger sequencing or SNaPshot methodologies) can be applied to age prediction in living individuals with similar accuracy. Hence, in both bisulfite sequencing and SNaPshot methodologies, we observed that the best APM developed for blood samples from living individuals, is not so accurate when applied to blood samples from deceased individuals. Because of this, we postulate that in cases of application to forensic contexts, the developed APMs should be applied to the same sample type and, more important, the health status of sample donor should be the same of the original model in order to achieve a more accurate age prediction. This

can be a challenge for forensic casework, in which, very often, the forensic specialist does not know the individual's state of life to select a specific model developed for living or deceased individuals. It may be advantageous to develop a single APM that combines information from the DNAm levels of blood samples from living and deceased individuals to be applied to forensic age estimate without information on the health status of the sample donor. When this is not possible, our data suggest that APMs developed using blood samples from deceased individuals could be applied to blood samples from living individuals, but not the contrary. We believe that this issue is an important point for development of new APMs or reproducibility of the existing APMs.

Finally, we should also consider that differences in sample size, DNAm markers, population age ranges, laboratory methodologies, or statistical techniques could influence accuracies across the different studies. In particular, bisulfite sequencing is a semi-quantitative method and thus may not be optimal for precise methylation analysis. Even though, using DNAm standards, the analysis for the best CpG site selected from each gene in living individuals (*ELOVL2* CpG6, *EDARADD* CpG3, *FHL2* CpG3 and *PDE4C* CpG2), in deceased individuals (*ELOVL2* CpG4, *EDARADD* CpG3, *FHL2* CpG2, *C1orf132* CpG1, and *PDE4C* CpG2) and in the overall sample set of blood samples (*ELOVL2* CpG6, *EDARADD* CpG3, *FHL2* CpG1 and *PDE4C* CpG2) indicates that the bisulfite sequencing method is accurate in terms of actual methylation *versus* expected methylation of known quantities of methylated to unmethylated DNA. Furthermore, the SNaPshot method demonstrated to be a promising method in forensic fields because of its capacity for multiplexing analysis, investigating simultaneously the DNAm levels across several specific CpGs previously reported as promising age-correlated markers. In our study, both methodologies shown to be efficient and economical alternative tools for rapid and efficient quantification of DNAm in blood samples. In agreement, when we

compare the age correlation captured by both methodologies for all the training sets, in general we observed similar values at the same CpGs (*ELOVL2*, Chr6:11044628; *FHL2*, Chr2:105399282; *C1orf132*, Chr1:207823681), supporting the accuracy of DNAm quantification of both methodologies for age predictions.

As conclusion, using a training set of 53 blood samples of living individuals, we developed a final APM with the highly age-associated CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* through the bisulfite sequencing methodology. The model revealed an accurate age estimation with a MAD from chronological age of 5.35 years. Our results are in concordance with previous studies using the pyrosequencing assay, being accurate for age estimations in forensic casework, at least in blood. Moreover, addressing age-dependency of multiple CpG sites in five genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* through bisulfite Sanger sequencing in blood samples of deceased individuals, the combination of the five most strong age-correlated markers from each gene in a final APM showed a MAD of 6.08 years that seems to be informative and accurate for age prediction purposes. This value revealed a similar accuracy comparing with the APM developed for blood samples from living individuals, showing reproducibility and applicability of bisulfite sequencing for age estimation in forensic contexts using blood. Bisulfite sequencing is simple, does not include complex procedures, being less time-consuming and less expensive, showing to have near equivalent accuracy to pyrosequencing (Jiang *et al.*, 2010; Parrish *et al.*, 2012). Moreover, bisulfite sequencing enabled us to assess the methylation level across multiple CpG sites allowing the possibility to choose several CpGs of interest.

In this study, we also validated and replicated in blood samples from living Portuguese individuals the SNaPshot assay of Jung *et al.* (2019) evaluating DNAm levels in the five specific CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59*

genes. Although using a different APM built with only three CpGs at *ELOVL2*, *FHL2* and *C1orf132* genes, similar high prediction accuracies were obtained (MAD = 4.25 years in Portuguese and MAD = 3.174 years in Koreans). Some differences in the methylation patterns were observed between the two populations (Koreans and Portuguese) suggesting the possibility that some effects in DNAm levels of age markers may be population specific. In deceased individuals, an independent APM with four CpGs in *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes was developed, showing a slightly lower prediction accuracy (MAD of 5.36 years) in relation to living individuals. Both MAD values obtained in APMs developed using SNaPshot methodology were similar, attesting the reproducibility of this methodology in blood samples from living and deceased individuals.

Our data revealed differences in DNAm levels between living and deceased individuals using both methodologies, suggesting that possible DNAm *postmortem* changes could alter the methylation status among specific loci. This is supported by differences in age correlation values, differences in the accuracy of the developed models and differences in selection of highest CpGs in each training set. Moreover, we observed that APMs developed in living individuals cannot be applied to dead people with identical accuracy, which reveals the necessity and usefulness of development of APMs specific not only for each type of sample but also considering the life and death status of the donor. Meanwhile, when this is not possible, our data suggest that APMs developed using blood from deceased individuals can be applied to living people. Furthermore, the development of APMs in the whole data set of samples from living and deceased individuals seems to be more suitable to be applied to forensic contexts when the status of sample donor is unknown.

Chapter 4. Results and discussion

B. DNA methylation age estimation in tooth samples

The obtained data from tooth samples from living and deceased individuals was used on the published original paper:

Correia Dias H, Corte Real F, Cunha E, Manco L. DNA methylation age estimation from human bone and teeth. Australian Journal of Forensic Sciences, 2020. Doi:10.1080/00450618.2020.1805011.

1. Introduction

Age estimation is one of the most relevant questions in forensic contexts, being necessary both for living and deceased individuals (Cunha *et al.*, 2009). Macroscopic and imagiological analyses followed by an appropriate mathematical approach are now standard procedures. But genetics and chemistry are also playing important roles in that respect (Zapico *et al.*, 2019). Odontological age estimation approaches have also played an important role in age estimation of living and deceased individuals and in non-adults and adults (Adserias-Garriga, 2019b). Despite, there is no standard method that can be applied for all the forensic scenarios, including living and deceased individuals and all the age ranges.

To date, few studies have considered teeth for DNA methylation (DNAm) analyses to predict age (Bekaert *et al.*, 2015a; Giuliani *et al.*, 2016; Márquez-Ruiz *et al.*, 2020). Giuliani *et al.* (2016) investigated methylation data at *ELOVL2*, *FHL2* and *PENK* loci by Maldi-Tof mass spectrometry in 21 modern teeth and proposed age prediction models (APMs) for cementum, dentin and pulp with a median absolute deviation between estimated and chronological ages of 2.45, 7.07 and 2.25 years, respectively. In concordance, Bekaert *et al.* (2015a) evaluating DNAm levels by pyrosequencing in 29 dentin samples from living individuals, reported a multiple quadratic regression model with seven CpGs located at *PDE4C*, *ELOVL2* and *EDARADD* genes explaining 74% of the variation in age with a mean absolute deviation (MAD) between estimated and chronological ages of 4.86 years. A recent study by Márquez-Ruiz *et al.* (2020) testing by bisulfite pyrosequencing the methylation levels in 65 tooth samples, obtained a significant positive age association for CpG sites at *ELOVL2* and *PDE4C* and developed an APM with nine CpGs from these two loci with a MAE of 5.08 years.

Considering these promising results, the main goal of the present study was to investigate DNAm information for age prediction purposes in tooth samples from identified Portuguese individuals using the bisulfite polymerase chain reaction (PCR) Sanger sequencing and multiplex SNaPshot methodologies. Both methods used herein are semi-quantitative but have shown to be efficient and economical alternative tools for rapid quantification of DNAm.

2. Materials and Methods

2.1. Sample collection

A set of 31 tooth samples (10 males, 21 females; aged 26-94 years old) was collected from living individuals (n = 23) in dentist offices, after informed consent, and from Bodies Donated to Science (BDS) (n = 8, Annex II) in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* before the embalming method of the body with Thiel (Eisma *et al.*, 2013). The collection of teeth before the embalming ensures the control of any possible influence related to the process of conservation.

The study protocol was approved by the *Instituto Nacional de Medicina Legal e Ciências Forenses* (INMLCF) and by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (n° 038-CE-2017).

2.2. DNA extraction, quantification and bisulfite conversion

Genomic DNA extraction from tooth samples was performed using a robot with *PrepFiler Express BTA™ Forensic DNA Extraction Kit* (Applied Biosystems, Foster City, CA). DNA Quantification was made using the real-time PCR-based *Quantifiler™ Human DNA Quantification Kit* (Applied Biosystems, Foster City, CA). These

procedures were made in INMLCF, according to standard guidelines, as previously described in *Chapter 3. Sample and design research*.

After extraction and quantification, genomic DNA was subjected to bisulfite conversion using *EZ DNA Methylation-Gold™ Kit* (Zymo Research, Irvine, USA) according to the instructions of manufacturer (previously described in *Chapter 3. Sample and design research*). Briefly, 20 µl of genomic DNA (in a total amount of 100 to 400 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 µl.

2.3. Polymerase chain reaction (PCR) and Sanger sequencing

After bisulfite conversion, the modified DNA samples were submitted to PCR for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany). PCR products were sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described in *Chapter 3. Sample and design research*.

2.4. SNaPshot assay

After bisulfite conversion, the modified DNA samples were submitted to a multiplex SNaPshot assay for five CpG sites at genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* with the primers and conditions previously described in Jung *et al.* (2019). Particular conditions for multiplex PCR amplification and multiplex SBE reactions were as previously described in *Chapter 3. Sample and design research*.

2.5. DNAm quantification

The methylation levels at each CpG site (0–1) was estimated by measuring the peak height of C (unconverted methylated DNA) and T (converted non-methylated DNA)

observed in the electropherograms through the formula $[C/C+T]$, as previously described in *Chapter 3. Sample and design research*.

2.6. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Independent analyses were made for the data obtained through Sanger sequencing and SNaPshot methodologies. Linear regression models were used to analyze relationships between DNAm levels and chronological age allowing the selection of the highest age-related CpGs and the development of APMs, as previously described in *Chapter 3. Sample and design research*. Briefly, using the simple linear regression coefficients from each significant age-correlated CpG site, we predicted age of individuals. The best combination of significant age-correlated CpGs selected from a stepwise regression analysis was used in a multiple regression approach to build the final multi-locus APM in each set of tooth samples.

The MAD between chronological and predicted ages and the root mean square error (RMSE) were calculated for the training set of tooth samples using the final APM developed in each methodology.

Validation of the final APMs for each set of tooth samples in both methodologies was performed by splitting the complete data set into two subsets (training and validation sets) and by a 3-fold cross validation as previously referred in *Chapter 3. Sample and design research*.

DNAm sex analysis was performed through comparison of two regression lines relating chronological age and DNAm levels of each gene at two levels (males/females)

of the categorical factor, using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA) in both methodologies.

3. Results

In the present study, we report on methylation levels of 43 CpG sites located at *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD* and *C1orf132* genes using the bisulfite Sanger sequencing methodology from tooth samples from living and deceased individuals. Moreover, using the multiplex SNaPshot assay reported by Jung *et al.* (2019), methylation data from five CpGs at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59* genes were obtained from teeth.

3.1. DNAm data obtained in teeth from living and deceased individuals using bisulfite Sanger sequencing

DNAm levels were evaluated in 31 tooth samples (23 from living and 8 from deceased individuals; 10 males, 21 females; aged 26-94 years old) through Sanger sequencing methodology.

Using DNAm standards, the accuracy of methylation levels obtained by bisulfite PCR sequencing in each best-selected CpG in the training set of tooth samples revealed a significant linear relationship to expected methylation levels (**Supplementary Figure S15**).

DNAm levels and sex

Two simple linear regression lines of methylation status and age between males and females showed no statistically significant difference in slope and intercept for the targets sites selected by Sanger sequencing in the training set of teeth (P -value >0.05 , **Table 4.26**). Thus, all the analyses were made ignoring sex differences.

Table 4.26: Comparison of two regression lines between males and females in tooth samples using data obtained from Sanger sequencing.

Marker	<i>P</i> -value	
	Intercept	Slope
<i>ELOVL2</i> CpG3	0.1621	0.6443
<i>FHL2</i> CpG4	0.1185	0.1696
<i>PDE4C</i> CpG1	0.2767	0.5262

Correlation between DNAm levels and chronological age

The DNAm levels in tooth samples showed no significant correlation with age for any CpG site at *EDARADD* and *C1orf132* genes (**Supplementary Table S8**). Thus, these loci were excluded for further analysis. The CpG sites at the remaining genes *ELOVL2*, *FHL2*, *PDE4C* showed positive but moderate or weak correlations between DNAm and age (**Supplementary Table S8**). One sample did not amplify for *FHL2* and four samples did not amplify for *PDE4C*. **Supplementary Figure S16** shows the correlation between age and DNAm levels of the best site in *ELOVL2*, *FHL2* and *PDE4C* genes.

Development of an age prediction model (APM)

Simple linear regression statistics testing the association between DNAm and chronological age revealed the strongest site *FHL2* CpG4 ($R = 0.658$, P -value = 0.000078), explaining 41.3% of the variation in age, followed by *PDE4C* CpG1 ($R =$

0.474, P -value = 0.013), explaining 19.3% of the variation in age, and *ELOVL2* CpG3 (R = 0.379, P -value = 0.036), explaining 11.4% of the variation in age (**Table 4.27; Supplementary Table S8**).

The predicted age of individuals was calculated through the simple linear regression coefficients for *FHL2* CpG4, *PDE4C* CpG1 and *ELOVL2* CpG3 allowing to obtain a MAD of 11.35 years, 14.58 years and 15.22 years, respectively (**Table 4.27; Figure 4.27; Supplementary Figure S17**).

Table 4.27: Simple linear regression statistics of the best age predictors in *FHL2*, *PDE4C* and *ELOVL2* genes to test the association between DNAm levels obtained by bisulfite sequencing and chronological age in tooth samples from living and deceased individuals.

Locus	CpG	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>FHL2</i>	CpG4	Chr2:105399297	30	0.658	0.433	0.413	14.57	0.000078	11.35
<i>PDE4C</i>	CpG1	Chr19:18233139	27	0.474	0.224	0.193	17.54	0.012588	14.58
<i>ELOVL2</i>	CpG3	Chr6:11044634	31	0.379	0.143	0.114	18.32	0.035738	15.22

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

We tested the age prediction multiple linear regression model using simultaneously the three best CpG sites, however only CpG4 at *FHL2* gene revealed a significant age correlation value (P -value = 0.002, **Supplementary Table S9**). Furthermore, we used the stepwise regression approach selecting the 12 significant age-correlated CpG sites located at *ELOVL2* (2 CpGs), *FHL2* (6 CpGs) and *PDE4C* (4 CpGs) but only the same *FHL2* CpG4 was chosen. Age prediction for each individual tooth estimated according to the individual regression coefficients for the high age-associated marker *FHL2* CpG4 was as follows: $(-114.989) + 239.863 \times \text{DNAm level } FHL2 \text{ CpG4}$ (**Table 4.28**). The predicted age captured from the 30 tooth samples allowed to obtain a moderate correlation between predicted and chronological ages of 0.589 (Spearman correlation) with a MAD from the chronological age of 11.35 years (RMSE = 14.08) (**Figure 4.27**). Correct predictions were 76.7% assuming that chronological and predicted ages match ± 15 years, according to the standard error of estimate calculated for the final APM (SE = 14.57) (**Table 4.27**).

Table 4.28: Statistical parameters obtained by simple linear regression for the *FHL2* CpG4, selected by stepwise regression approach in tooth samples.

Marker	Coefficient	P -value
(Intercept)	-114.989	0.005
<i>FHL2</i> CpG4	239.863	0.000

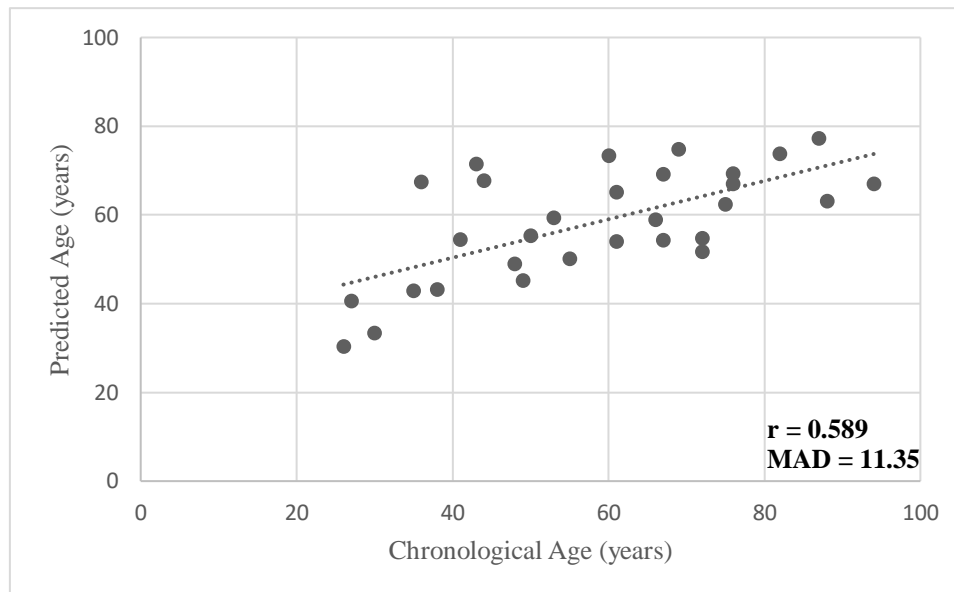


Figure 4.27: Predicted age *versus* chronological age using best marker *FHL2* CpG4 in teeth from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Validation of the simple APM developed in living and deceased individuals

The simple APM developed for tooth samples with *FHL2* CpG4 was tested by a 3-fold cross validation allowing an averaged MAD from the chronological age of 12.22 years (RMSE = 12.36) from the three independent validation sets. The additional validation by splitting the overall training set into two sets of 16 and 14 samples (training and test sets) allowed to obtain a MAD of 10.40 years (RMSE = 13.20) in the training set and 11.93 years (RMSE = 15.98) in the test set.

3.2. DNAm data obtained in teeth from living and deceased individuals using SNaPshot methodology

From the 31 tooth samples from living and deceased individuals evaluated using the multiplex methylation SNaPshot assay of Jung *et al.* (2019), 24 samples (16 from

living and 8 from deceased individuals; 8 males, 16 females; aged 27-88 years old) were successfully PCR-amplified and the obtained data were used for subsequent analysis.

DNAm levels and sex

DNAm sex analysis in the training set of teeth analyzed by SNaPshot methodology revealed no statistically significant difference in slope and intercept for sites located at *ELOVL2*, *KLF14* and *TRIM59* loci (P -value >0.05 , **Table 4.29**). Thus, all the analyses were made ignoring sex differences.

Table 4.29: Comparison of two regression lines between males and females in tooth samples using data obtained from the SNaPshot methodology.

Marker	<i>P</i> -value	
	Intercept	Slope
<i>ELOVL2</i>	0.6102	0.0922
<i>KLF14</i>	0.4980	0.2465
<i>TRIM59</i>	0.9121	0.0714

Correlation between DNAm levels and chronological age

From the 24 tooth samples with successful amplifications using the multiplex methylation SNaPshot assay, *ELOVL2*, *KLF14* and *FHL2* genes were not successfully amplified for one sample. Positive and significant age correlations were observed for *ELOVL2*, *KLF14* and *TRIM59* genes in the training set sample (**Table 4.30; Supplementary Figure S18**). The CpG sites in *FHL2* and *C1orf132/MIR29B2C* genes showed a lower or no significant age correlation (**Table 4.30**) and were excluded for further analysis.

Development of an age prediction model (APM)

DNAm levels in *KLF14* locus showed the strongest age correlation ($R = 0.728$, P -value = 0.000084), explaining 50.7% of the variation in age, following by *ELOVL2* ($R = 0.685$, P -value = 0.000311), explaining 44.4% of the variation in age, and *TRIM59* ($R = 0.665$, P -value = 0.000389), explaining 41.7% of the variation in age (**Table 4.30**). Using these sites in simple linear regression models the calculated MAD values from chronological age were 9.68 years for *KLF14*, 11.27 years for *ELOVL2* and 11.51 years for *TRIM59* (**Supplementary Figure S19**).

B. DNA methylation age estimation in tooth samples

Table 4.30: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using SNaPshot assay in teeth from living and deceased individuals.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>ELOVL2</i>	Chr6:11044628	23	0.685	0.469	0.444	13.33	0.000311	11.27
<i>FHL2</i>	Chr2:105399282	23	0.331	0.110	0.067	17.43	0.122700	-
<i>KLF14</i>	Chr7:130734355	23	0.728	0.529	0.507	12.56	0.000084	9.68
<i>C1orf132</i>	Chr1:207823681	24	-0.080	0.006	-0.039	19.07	0.709684	-
<i>TRIM59</i>	Chr3:160450189	24	0.665	0.443	0.417	14.28	0.000389	11.51
<i>Multiple linear regression</i>								
APM (<i>ELOVL2</i> and <i>KLF14</i>)		23	0.886	0.785	0.764	12.56	2.09×10^{-7}	7.07

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

The training set of 23 teeth was used to develop a multiple linear regression model using simultaneously the methylation information of the three individual significant age-correlated CpG sites located at *KLF14*, *ELOVL2* and *TRIM59* genes. However, the CpG site at *TRIM59* gene showed non-significant age correlation value in the multiple model, which could reveal signs of multicollinearity between variables (**Supplementary Table S10**). In concordance, the stepwise regression approach allowed to select a final dual-locus model with the CpG sites at *ELOVL2* and *KLF14*. The multiple linear regression using *ELOVL2* and *KLF14* DNAm data allowed us to obtain a strong age correlation value ($R = 0.886$), highly significant ($P\text{-value} = 2.09 \times 10^{-7}$), explaining 76.4% of the variation in age (**Table 4.30**). Predicting age of each individual through the formula: $11.519 + 106.261 \times \text{DNAm level } ELOVL2 + 291.877 \times \text{DNAm level } KLF14$ (**Table 4.31**), allowed to estimate age with a correlation between predicted and chronological ages of 0.883 (Spearman correlation), and a MAD from chronological age of 7.07 years (RMSE = 8.11) (**Figure 4.28**). Correct predictions were 95.7% assuming that chronological and predicted ages match ± 13 years, according to the standard error of estimate calculated for the final APM (SE =12.56) (**Table 4.30**).

Table 4.31: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *ELOVL2* and *KLF14* selected by stepwise regression approach, in tooth samples.

Marker	Coefficient	P-value
(Intercept)	11.519	0.061
<i>ELOVL2</i>	106.261	0.000
<i>KLF14</i>	291.877	0.000



Figure 4.28: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the two CpGs located at *ELOVL2* and *KLF14* genes in teeth from living and deceased individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Validation of the multi-locus APM developed in living and deceased individuals

The 3-fold cross validation made to test the accuracy of the model, allowed to estimate an averaged MAD from the chronological age for the three independent validation sets of 7.33 years (RMSE = 7.48), very close to the MAD of 7.07 years from the total training data set. A validation by splitting the sample into two sets of 12 and 11 samples (training and validation sets) allowed to obtain an independent MAD value for the training set of 7.35 years (RMSE = 8.59). The model was applied to the validation set allowing to obtain a MAD of 6.34 years (RMSE = 7.25).

3.3. Comparison between methodologies

The age correlation values obtained in tooth samples using Sanger sequencing and SNaPshot methodologies revealed similar moderate age correlations for *FHL2* ($R = 0.451$ vs. $R = 0.331$, respectively) and a similar negligible age correlation for *C1orf132* ($R = -$

0.123 vs. $R = -0.080$, respectively). *ELOVL2* marker showed a higher age correlation value through SNaPshot ($R = 0.685$) comparing with Sanger sequencing ($R = 0.237$) (Table 4.32).

Table 4.32: Comparison of age-correlated values obtained in tooth samples from living and deceased Portuguese individuals through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Teeth (26-94 years)		Teeth (27-88 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.237	0.056	0.685	0.469
<i>C1orf132</i> Chr1:207823681	-0.123	0.015	-0.080	0.006
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.451	0.204	0.331	0.110

Abbreviations: R, Pearson correlation coefficient.

Differences between predicted and chronological ages with aging were not evaluated for both methodologies, because we have a limited number of 31 tooth samples that is not representative of all age groups.

4. Discussion

Nowadays, the increasing growth of epigenetics in forensic contexts led to the development of DNAm models to predict age based in different sample types. However, only few studies have focused in DNAm age prediction based on teeth (Bekaert *et al.*, 2015a; Giuliani *et al.*, 2016; Márquez-Ruiz *et al.*, 2020). Teeth are a valuable source of DNA in forensics, due to its high resistance to adverse conditions (including *postmortem*

DNA decay). Very often, this tissue has better quality than bone samples (Higgins and Austin, 2013).

In the present study, we analyzed 31 tooth samples of Portuguese ancestry through the bisulfite Sanger sequencing method to obtain methylation information of several CpGs located at *ELOVL2* (9 CpGs), *FHL2* (12 CpGs), *EDARADD* (4 CpGs), *PDE4C* (12 CpGs) and *C1orf132* (6 CpGs) genes, repeatedly reported as age-associated in blood (Weidner *et al.*, 2014; Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015b; Thong *et al.*, 2017; Cho *et al.*, 2017; Duanay *et al.*, 2019; Pfeifer *et al.*, 2020). Moreover, we used the multiplex methylation SNaPshot assay of Jung *et al.* (2019) in 24 tooth samples to analyze five CpG sites located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59*. SNaPshot methodology seems promising in forensic fields because of its capacity for multiplexing analysis.

The Sanger sequencing methodology did not allow to develop a multi-locus APM in tooth samples, since only lower and moderate correlation values between age and DNAm were obtained. This suggests that this method could have limited usefulness for forensic age estimation using teeth. Meanwhile, a simple linear regression model using the high age-correlated *FHL2* CpG4 revealed a model accuracy of 11.35 years, which is a higher MAD value. Testing this APM by a 3-fold cross validation (MAD = 12.22 years) and the validation by splitting the overall training set into two sets (MAD = 10.40 years in the training set and MAD = 11.93 years in the test set) showed similar values, suggesting the accuracy of the APM.

Using the SNaPshot method, a moderate accurate multiple model combining CpGs at *ELOVL2* and *KLF14* was obtained (MAD of 7.07 years). Testing the developed model by the 3-fold cross validation (MAD = 7.33 years) and through the validation by splitting the sample into two sets (MAD = 7.35 years for the training set and 6.34 years

for the validation set) showed MAD values very close to the MAD of 7.07 years obtained from the total training data set. This suggests that accurate results can be obtained when testing the developed models.

Regarding previous studies using teeth, Bekaert *et al.* (2015a), evaluating DNAm levels by pyrosequencing in 29 dentin samples of living individuals, reported a multiple quadratic regression model developed with seven CpGs located at *PDE4C*, *ELOVL2* and *EDARADD* genes explaining 74% of the variation in age with a MAD of 4.86 years. The model included the *ELOVL2* CpG site (Chr6:11044628) selected in our study using SNaPshot. Giuliani *et al.* (2016) using five CpGs located at *ELOVL2*, *FHL2* and *PENK* genes, addressed by Maldi-Tof mass spectrometry in 21 teeth extracted from living individuals with age ranging from 17 to 77 years, obtained a median absolute deviation of 7.07 years in dentin samples. A recent study by Márquez-Ruiz *et al.* (2020) testing methylation levels of specific CpG sites located in the *ELOVL2* and *PDE4C* genes by bisulfite pyrosequencing in 65 tooth samples from individuals aged 15-85 years old, developed an APM with nine CpG sites showing a MAE of 5.08 years. All these studies showed the usefulness of *ELOVL2* in development of epigenetic clocks using tooth samples. In agreement, our dual-model obtained with the SNaPshot method included a CpG from this locus. Similar to Bekaert *et al.* (2015a) and Márquez-Ruiz *et al.* (2020), our study showed non-significant sex DNAm differences, using Sanger sequencing and SNaPshot methodologies.

The age correlation in tooth samples using both methodologies revealed similar values for *FHL2* and *C1orf132* markers. However, for CpG from *ELOVL2* a slight difference was observed between both methodologies emphasizing the need of additional studies with a more representative sample set of tooth samples.

The present study has some limitations, being the sample size the major drawback. Moreover, despite we used 31 teeth not all of them were successful amplified for all genes. Larger sample sets could have higher statistical power and could be more representative of DNAm changes according to different age range and in different tissue types. We should consider also that different CpGs and/or genes were addressed in both methodologies, consequently this could influence the accuracy of the developed models.

In conclusion, considering that to date only few reports used teeth in development of models for forensic age estimation, we evaluated DNAm levels from tooth samples using the bisulfite Sanger sequencing and SNaPshot methodologies. Our study allowed to develop a dual-locus APM with CpGs located at *ELOVL2* and *KLF14* genes through SNaPshot method, exhibiting a moderate age prediction accuracy, allowing to obtain a MAD from chronological age of 7.07 years. Our results suggest that bisulfite PCR sequencing could have limited efficacy using tooth samples, being a complementary methodology for evaluating DNAm changes and creating age estimation models in this type of tissue.

Chapter 4. Results and discussion

C. DNA methylation age estimation in fresh bone samples

The obtained data from bone samples collected during autopsies was used on the published original paper:

Correia Dias H, Corte Real F, Cunha E, Manco L. DNA methylation age estimation from human bone and teeth. Australian Journal of Forensic Sciences 2020. Doi:10.1080/00450618.2020.1805011.

1. Introduction

Bones can be one of the last evidences of an individual and provide an important source for DNA in many cases of forensic identification. Thus, bones can be one of the most informative samples for forensic purposes (Meissner and Ritz-Timme, 2010). Estimating an individual's age based on skeletal indicators is not new in forensic casework, and many age estimation methods based on anthropological features were implemented (Adserias-Garriga and Wilson-Taylor, 2019; Cunningham, 2019); however, these methods can be affected by some factors as characteristics of the reference sample and subjectivity (Cunha *et al.*, 2009; Franklin, 2010; Nakhaeizadeh *et al.*, 2014). Additionally, molecular or chemical methodologies can be a promising alternative for age estimation (Zapico *et al.*, 2019), but can also suffer from practical difficulties related to standardization of methods or implementation in many forensic contexts (Meissner and Ritz-Timme, 2010).

DNA methylation (DNAm) based models for age estimation are increasingly accepted as a promising tool to be applied in future forensic contexts. However, to date, few studies have assessed to DNAm levels in bone samples for age estimation purposes (Horvath *et al.*, 2015, 2018; Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020). Horvath *et al.* (2015) applied a previously developed model with 353 CpGs (Horvath, 2013) to 48 trabecular bone samples, obtaining a correlation value of 0.88 between predicted and chronological ages. Later, Horvath *et al.* (2018) developed the “skin & blood clock” focused on 391 CpGs analyzed in several tissues, including blood, skin, buccal swabs and saliva. When applying the “skin & blood clock” to trabecular bone samples, the authors obtained a correlation between predicted and chronological ages of 0.82 (Horvath *et al.* 2018). Using massive parallel sequencing, Naue *et al.* (2018)

obtained in fresh bone samples collected during autopsy from 29 individuals a moderate correlation between DNAm and age for CpGs at *ELOVL2*, *KLF14* and *TRIM59* genes (0.58, 0.51 and 0.61, respectively). Gopalan *et al.* (2019) using genome-wide DNAm data from 155 bone samples developed a powerful “37 bone clock CpGs” for age prediction based on CpG sites from *TRIM59*, *ELOVL2* and *KLF14* genes, among others. They obtained a root mean square error (RMSE) of 4.9 years. They investigated also forensic samples with 2-3 years of natural decomposition, however these samples were excluded from analysis. More recently, Lee *et al.* (2020) observed for *TRIM59* and *ELOVL2* CpGs an age correlation value of 0.434 and 0.415, respectively, when evaluated 30 bones using the SNaPshot assay proposed by Jung *et al.* (2019). In addition, Lee and collaborators applied the “skin & blood clock” previously developed by Horvath *et al.* (2018) in 12 bones using the Infinium HumanMethylation EPIC BeadChip array, observing strong age correlation values ($r > 0.70$) for 38 CpGs, including for CpGs *FHL2* (cg06639320, cg22454769), *EDARADD* (cg09809672), *ELOVL2* (cg21572722, cg16867657), *KLF14* (cg20426994) and *TRIM59* (cg07553761). Predicting age using the 12 bones allowed to obtain a mean absolute deviation (MAD) from chronological age of 6.4 years ($r = 0.964$) (Lee *et al.*, 2020).

Considering the scarcity of studies in development of age prediction models (APMs) based on DNAm analysis using bones, the development of specific models for this tissue type is required, as well as the discovery of the best age-related markers. To address this issue, the present study aimed to develop APMs for bone samples from Portuguese individuals, using the bisulfite Sanger sequencing and SNaPshot methodologies. The previously known and validated age-associated genes *ELOVL2*, *EDARADD*, *PDE4C*, *FHL2*, *C1orf132/MIR29B2CHG*, *TRIM59* and *KLF14* were used in the analyses.

2. Materials and Methods

2.1. Sample collection

A set of 31 bone samples (26 males, 5 females; aged 26-81 years old) was collected from identified deceased individuals during autopsies in *Serviço de Patologia Forense das Delegações do Centro e Sul*, after consulting RENNDA (*Registo Nacional de Não Dadores*). All these bones were collected within five days after death.

Moreover, a training set of 22 bones (10 males, 12 females; aged 49-93 years old) collected from identified Bodies Donated to Science (BDS) was included. Bones from BDS were collected in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* and in the *Faculdade de Medicina da Universidade de Coimbra*, after the embalming method with Thiel (Eisma *et al.*, 2013). The training set of bones from BDS has different *postmortem* intervals (Annex II).

The study protocol was approved by the *Instituto Nacional de Medicina Legal e Ciências Forenses* (INMLCF) and by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (n° 038-CE-2017).

2.2. DNA extraction, quantification and bisulfite conversion

The pre-treatment of fresh bone samples, genomic DNA extraction using *PrepFiler Express BTA™ Forensic DNA Extraction Kit* (Applied Biosystems, Foster City, CA) and DNA Quantification with the real-time PCR-based *Quantifiler™ Human DNA Quantification Kit* (Applied Biosystems, Foster City, CA) were made in INMLCF, as described in *Chapter 3. Sample and design research*. Then, genomic DNA was subjected to bisulfite conversion using the *EZ DNA Methylation-Gold™ Kit* (Zymo

Research, Irvine, USA) according to the instructions of manufacturer. Briefly, 20 µl of genomic DNA (in a total amount of 100 to 400 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 µl. Specific details were previously described in *Chapter 3. Sample and design research*.

2.3. Polymerase chain reaction (PCR) and Sanger sequencing

After bisulfite conversion, the modified DNA samples were submitted to PCR for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described in *Chapter 3. Sample and design research*.

2.4. SNaPshot assay

After bisulfite conversion, the modified DNA samples were submitted to a multiplex SNaPshot assay for five CpG sites at genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* with the primers and conditions described in Jung *et al.* (2019). Particular conditions for multiplex PCR amplification and multiplex SBE reactions were as previously described in *Chapter 3. Sample and design research*.

2.5. DNAm quantification

The methylation levels at each CpG site (0–1) was estimated by measuring the peak height of C (unconverted methylated DNA) and T (converted non-methylated DNA) observed in the electropherograms through the formula $[C/C+T]$, as previously described in *Chapter 3. Sample and design research*.

2.6. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Independent analyses were made for each training set of bones using data obtained through bisulfite sequencing and SNaPshot methodologies. Linear regression models were used to analyze relationships between methylation levels and chronological age as previously described in *Chapter 3. Sample and design research*. Briefly, using the simple linear regression coefficients from each significant age-correlated CpG site, we predicted age of individuals for each individual gene. The best combination of significant age-correlated CpGs, selected from a stepwise regression analysis was used in a multiple regression approach to build the final multi-locus APM in each set of bones.

The MAD between chronological and predicted ages and the root mean square error (RMSE) were calculated for each training set of bone samples using the final APM developed in each methodology.

Validation of the final APMs for each set of bones in both methodologies was performed by splitting the complete data set into two subsets (training and validation sets), and by a 3-fold cross validation, as previously referred in *Chapter 3. Sample and design research*.

Sex DNAm analysis was made using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA), as reported in *Chapter 3. Sample and design research* for the training set of bones from BDS in both methodologies.

The effect of PMI (*postmortem* interval) was evaluated in six pairs of individuals from the BDS, with identical or similar chronological age, calculating the mean of the differences between DNAm levels obtained for all the CpGs within each gene.

3. Results

In the present study, we evaluated the DNAm levels of several CpG sites located at genes *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *KLF14* and *TRIM59* in bones from autopsies and from BDS using the bisulfite Sanger sequencing and the SNaPshot methodologies.

Using three known quantities of DNAm standards (0%, 50% and 100% methylation levels), the methylation data obtained by bisulfite sequencing revealed a linear relationship between the observed and the expected DNAm levels of the best age-correlated CpG site in each training set of samples (bones from autopsies and from BDS) (**Supplementary Figure S20**).

3.1. Age estimation in bones collected during autopsies

3.1.1. DNAm data obtained in bones from autopsies using bisulfite Sanger sequencing

DNAm levels of 43 CpG sites located at *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD* and *C1orf132* genes were evaluated by the bisulfite Sanger sequencing methodology from 29 bones from autopsies (4 females, 25 males; aged 26-80 years old).

Correlation between DNAm levels and chronological age

Bone samples revealed positive correlations for all the CpG sites located at genes *ELOVL2*, *FHL2* and *PDE4C* and negative correlations for the CpG sites on the remaining

two genes *EDARADD* and *C1orf132* (**Supplementary Table S11**). Plots with the best CpG in each gene are present in **Supplementary Figure S21**.

Simple linear regression showed strong and significant age correlation values ($0.70 < R < 0.90$) for almost all CpGs located at *ELOVL2* and for the three first CpG sites at *C1orf132* gene. The remaining markers showed negligible, weak or moderate age correlations (**Supplementary Table S11**). One sample did not amplify for the *PDE4C* gene.

Development of an age prediction model (APM)

Testing the association between chronological age and DNAm levels for the individual 43 CpG sites, the strongest age-correlated value was obtained for the *ELOVL2* CpG6 ($R = 0.852$; $P\text{-value} = 4.64 \times 10^{-9}$), explaining 71.5% of the variation in age, following by *C1orf132* CpG1 ($R = -0.834$; $P\text{-value} = 1.93 \times 10^{-8}$), explaining 68.4% of the variation in age (**Table 4.33; Supplementary Table S11**). For the remaining genes, the best markers showed moderate age correlation values: *FHL2* CpG1 ($R = 0.692$, $P\text{-value} = 0.000032$), explaining 46.0% of the variation in age; *PDE4C* CpG2 ($R = 0.690$; $P\text{-value} = 0.000049$), explaining 45.6% of the variation in age; and *EDARADD* CpG3 ($R = -0.561$; $P\text{-value} = 0.001564$), explaining 28.9% of the variation in age (**Table 4.33; Supplementary Table S11**).

Simple linear regression APMs based on these individual CpGs revealed MAD values from chronological age of 5.73 years for *ELOVL2* CpG6, 6.35 years for *C1orf132* CpG1, 8.39 years for *FHL2* CpG1, 8.07 years for *PDE4C* CpG2 and 9.23 years for *EDARADD* CpG3 (**Table 4.33; Supplementary Figure S22**).

Table 4.33: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in bone samples from autopsies.

Locus	CpG	Location	N	R	R ²	Correc ted R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	CpG6	Chr6:11044644	29	0.852	0.725	0.715	7.24	4.64×10^{-9}	5.73
<i>C1orf132</i>	CpG1	Chr1:207823681	29	-0.834	0.695	0.684	7.63	1.93×10^{-8}	6.35
<i>FHL2</i>	CpG1	Chr2:105399282	29	0.692	0.479	0.460	9.97	0.000032	8.39
<i>PDE4C</i>	CpG2	Chr19:18233133	28	0.690	0.476	0.456	10.19	0.000049	8.07
<i>EDARADD</i>	CpG3	Chr1:236394382	29	-0.561	0.314	0.289	11.45	0.001564	9.23
<i>Multiple linear regression</i>									
APM (<i>ELOVL2</i> CpG5, <i>ELOVL2</i> CpG7, <i>ELOVL2</i> CpG6, <i>C1orf132</i> CpG1, <i>EDARADD</i> CpG3 and <i>EDARADD</i> CpG4)			29	0.970	0.941	0.925	3.71	2.097×10^{-12}	2.56

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

We selected all the 29 CpG sites which showed individual significant association with age located at *ELOVL2* (9 CpGs), *FHL2* (7 CpGs), *EDARADD* (3 CpGs), *PDE4C* (5 CpGs) and *C1orf132* (5 CpGs) and applied the stepwise linear regression analysis to select other relevant variables to be used in a final multi-locus APM (**Table 4.34**).

Table 4.34: Statistical parameters obtained in a multiple regression model with the six CpGs in genes *ELOVL2*, *C1orf132* and *EDARADD* selected by stepwise regression approach, in bone samples from autopsies.

Marker	Coefficient	P-value
(Intercept)	129.912	0.004
<i>EDARADD</i> CpG3	-136.346	0.000
<i>EDARADD</i> CpG4	62.928	0.013
<i>C1orf132</i> CpG1	-66.051	0.005
<i>ELOVL2</i> CpG5	67.573	0.000
<i>ELOVL2</i> CpG6	144.915	0.001
<i>ELOVL2</i> CpG7	-137.429	0.004

The stepwise regression selected six CpGs located at *ELOVL2* (CpG5, CpG6, CpG7), *EDARADD* (CpG3, CpG4) and *C1orf132* (CpG1) and revealed in the multiple regression analysis a high age correlation coefficient ($R = 0.970$), explaining 92.5% of the variation in age, highly significant ($P\text{-value} = 2.097 \times 10^{-12}$) (**Table 4.33**). Age prediction for each individual bone sample was estimated according to the equation developed with the regression coefficients as present in **Table 4.34**: $129.912 - 66.051 \times \text{DNAm level } C1orf132 \text{ CpG1} - 136.346 \times \text{DNAm level } EDARADD \text{ CpG3} + 62.928 \times \text{DNAm level } EDARADD \text{ CpG4} + 67.573 \times \text{DNAm level } ELOVL2 \text{ CpG5} + 144.915 \times \text{DNAm level } ELOVL2 \text{ CpG6} - 137.429 \times \text{DNAm level } ELOVL2 \text{ CpG7}$. The developed APM enabled us to estimate age with a correlation between predicted and chronological ages of 0.957 (Spearman correlation) and a MAD from chronological age of 2.56 years

(RMSE = 3.24) (**Figure 4.29**). Correct predictions were 79.3% according to the standard error of estimate calculated for the final APM (SE = 3.71).

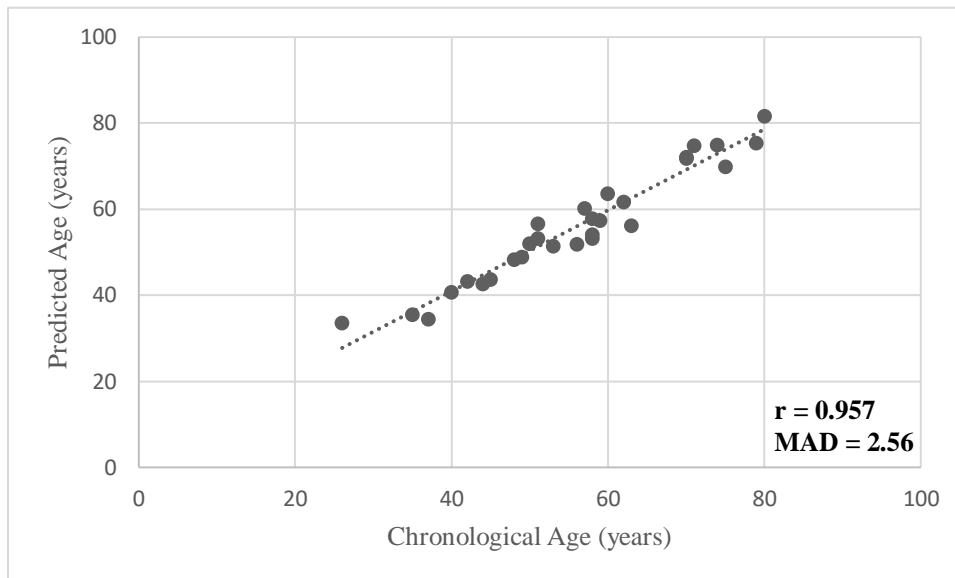


Figure 4.29: Predicted age *versus* chronological age using the six markers *ELOVL2* CpG5, *ELOVL2* CpG6, *ELOVL2* CpG7, *EDARADD* CpG3, *EDARADD* CpG4 and *C1orf132* CpG1 in bones from autopsies. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Validation of the multi-locus APM developed in deceased individuals

The 3-fold cross validation to test the accuracy of the model, showed an averaged MAD from the chronological age for the three independent validation sets of 3.77 years (RMSE = 3.93). This value is very close to the MAD of 2.56 years of the whole training set.

Additionally, a second validation method by splitting the sample (29 bones) into two sets of 16 and 13 samples (training and validation sets) revealed an independent MAD value for the training set of 2.03 years (RMSE = 2.70). The multiple linear regression equation developed in the training set of 16 samples was applied to the validation set allowing to obtain a MAD of 3.15 years (RMSE = 4.08).

3.1.2. DNAm data obtained in bones from autopsies using SNaPshot methodology

Using the multiplex SNaPshot assay reported by Jung *et al.* (2019), methylation data from the five CpG sites at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were obtained from 31 bones from autopsies (26 females, 5 males; aged 26-81 years old).

Correlation between DNAm levels and chronological age

Positive correlations with age were observed for CpGs at genes *ELOVL2*, *FHL2*, *KLF14* and *TRIM59*, and a negative correlation was obtained for *C1orf132* locus (**Supplementary Figure S23**).

Development of an age prediction model (APM)

Among the five markers, the CpG site in the *FHL2* locus showed the strongest age correlation ($R = 0.708$, $P\text{-value} = 0.000008$), explaining 48.4% of the variation in age, followed by *TRIM59* ($R = 0.633$, $P\text{-value} = 0.000129$), explaining 38.1% of the variation in age, *ELOVL2* ($R = 0.619$, $P\text{-value} = 0.000202$), explaining 36.3% of the variation in age, *KLF14* ($R = 0.540$, $P\text{-value} = 0.001708$), explaining 26.7% of the variation in age, and *C1orf132* ($R = -0.507$, $P\text{-value} = 0.003640$), explaining 23.1% of the variation in age (**Table 4.35; Supplementary Figure S23**). Simple APMs using each CpG site revealed MAD values from the chronological age of 7.95 years for *FHL2*, 8.50 years for *ELOVL2*, 8.86 years for *TRIM59*, 9.29 years for *KLF14* and 9.70 years for *C1orf132* genes (**Supplementary Figure S24**).

Table 4.35: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* loci using SNaPshot assay in bones from autopsies.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>FHL2</i>	Chr2:105399282	31	0.708	0.501	0.484	10.26	0.000008	7.95
<i>TRIM59</i>	Chr3:160450189	31	0.633	0.402	0.381	11.23	0.000129	8.86
<i>ELOVL2</i>	Chr6:11044628	31	0.619	0.384	0.363	11.40	0.000202	8.50
<i>KLF14</i>	Chr7:130734355	31	0.540	0.292	0.267	12.23	0.001708	9.29
<i>C1orf132</i>	Chr1:207823681	31	-0.507	0.257	0.231	12.53	0.003640	9.70
<i>Multiple linear regression</i>								
APM (<i>FHL2</i> and <i>KLF14</i>)		31	0.777	0.604	0.576	9.30	0.000002	7.18

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Multiple linear regression analysis with simultaneously all the five predictor variables revealed non-significant p-values for few predictors (**Supplementary Table S12**). In concordance, fitting the stepwise regression analysis by using simultaneously the five CpG sites, only CpG sites at *FHL2* and *KLF14* genes were chosen. The final multiple dual-locus model built with CpG sites at *FHL2* and *KLF14* showed a moderate age correlation value ($R = 0.777$; $P\text{-value} = 0.000002$), explaining 57.6% of the variation in age (**Table 4.35**). The developed formula obtained with the regression coefficients to predict age was as follows: $15.727 + 105.392 \times \text{DNAm level } FHL2 + 154.672 \times \text{DNAm level } KLF14$ (**Table 4.36**).

Table 4.36: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *FHL2* and *KLF14* selected by stepwise regression approach, in bone samples from autopsies.

Marker	Coefficient	P-value
(Intercept)	15.727	0.027
<i>FHL2</i>	105.392	0.000
<i>KLF14</i>	154.672	0.012

The two-locus model showed a strong correlation of 0.746 between predicted and chronological ages (Spearman correlation value), with a MAD from chronological age of 7.18 years (RMSE = 8.84) (**Figure 4.30**). Correct predictions were 71% assuming that chronological and predicted ages match ± 9 years, according to the standard error of estimate calculated for the final APM (SE = 9.30).

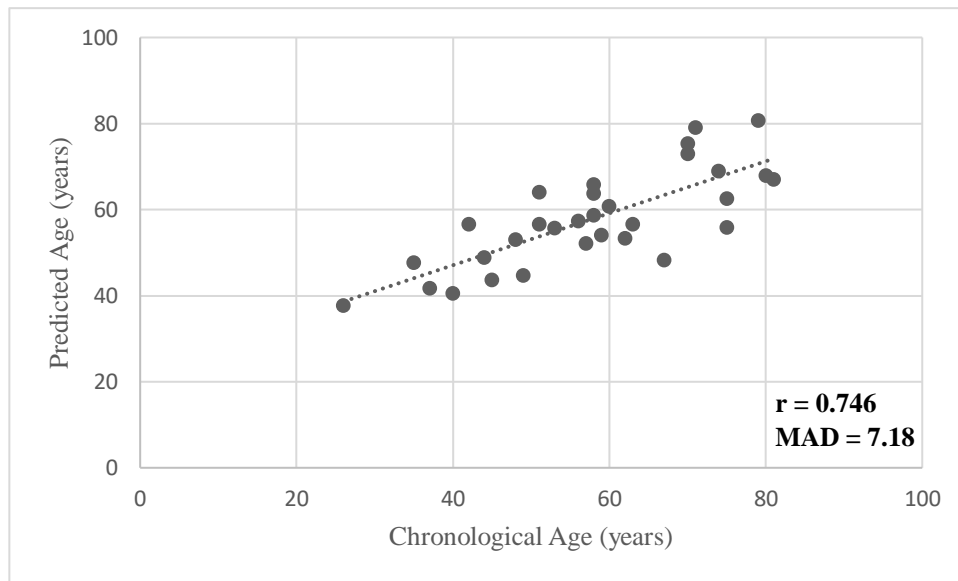


Figure 4.30: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the two CpGs located at *FHL2* and *KLF14* genes in bones from autopsies. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Validation of the multi-locus APM developed in deceased individuals

The 3-fold cross validation to test the accuracy of the model, showed an averaged MAD from the chronological age for the three independent validation sets of 7.84 years (RMSE = 7.86), a value very close to the MAD of 7.18 of the whole training set. The validation by splitting the sample into two sets of 16 and 15 samples each (training and validation sets) allowed to obtain an independent MAD from the chronological age for the training set of 5.40 years (RMSE = 6.78) and for the validation set of 9.35 years (RMSE = 12.07).

3.1.3. Comparison between methodologies

Comparing the age correlation value obtained in bones from autopsies using Sanger sequencing and SNaPshot, we observed similar age association for *ELOVL2* Chr6:11044628 ($R = 0.736$ vs. $R = 0.619$, respectively) and *FHL2* Chr2:105399282 ($R = 0.692$ vs. $R = 0.708$, respectively). The *C1orf132* Chr1:207823681 revealed a stronger

age-correlated value for Sanger compared to SNaPshot ($R = -0.834$ vs. $R = -0.507$, respectively) (**Table 4.37**).

Table 4.37: Comparison of age-correlated values obtained in bones from autopsies through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Bone deceased (26-80 years)		Bone deceased (26-81 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.736	0.542	0.619	0.384
<i>C1orf132</i> Chr1:207823681	-0.834	0.695	-0.507	0.257
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.692	0.479	0.708	0.501

Abbreviations: R, Pearson correlation coefficient.

3.2. Age estimation in bones collected from BDS

3.2.1. DNAm data obtained in bones from BDS using bisulfite Sanger sequencing

A total of 43 CpGs located at genes *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD* and *C1orf132* was evaluated in 22 bones from BDS (10 males, 12 females; aged 49-93 years old) through Sanger sequencing methodology. One sample was not analyzed for *PDE4C* due to complete PCR failure.

DNAm sex analysis revealed no statistically significant difference between males and females for the selected target sites (P -value >0.05 , **Table 4.38**).

Table 4.38: Comparison of two regression lines between males and females in bones from BDS using data obtained from Sanger sequencing.

Marker	P-value	
	Intercept	Slope
<i>FHL2</i> CpG3	0.1532	0.5363
<i>FHL2</i> CpG2	0.8553	0.5905
<i>EDARADD</i> CpG3	0.1366	0.8291
<i>PDE4C</i> CpG3	0.9887	0.7717

Correlation between DNAm levels and chronological age

Evaluating the association between DNAm levels and chronological age, we observed positive age correlations for CpGs from *FHL2* and *PDE4C* genes, while a negative age correlation was observed for *EDARADD* gene (**Supplementary Table S13**). Plots with the best CpG in these three genes are present in **Supplementary Figure S25**. The *ELOVL2* and *C1orf132* genes showed negligible or very weak age correlation values and were excluded from further analysis (**Supplementary Table S13**).

Development of an age prediction model (APM)

The strongest age correlation was observed for *PDE4C* CpG3 ($R = 0.771$, P -value = 0.000044), explaining 57.2% of the variation in age, followed by *FHL2* CpG3 ($R = 0.655$, P -value = 0.000945), explaining 40.0% of the variation in age, and *EDARADD* CpG3 ($R = -0.430$, P -value = 0.0459), explaining only 14.4% of the variation in age (**Table 4.39; Supplementary Table S13**). Simple linear regression analysis for each CpG showed MAD values from chronological age of 5.53 years for *PDE4C* CpG3, 7.12 years for *FHL2* CpG3 and 8.57 years for *EDARADD* CpG3 (**Table 4.39; Supplementary Figure S26**).

We tested the age prediction multiple linear regression model using simultaneously these three CpG sites, however no significant age correlation values were obtained for *EDARADD* CpG3 (P -value = 0.283) and *FHL2* CpG3 (P -value = 0.505)

(Supplementary Table S14). In concordance, through stepwise regression approach, only *PDE4C* CpG3 was chosen.

Trying to address other relevant sites for the development of a more accurate APM, we use the stepwise regression approach selecting the 10 significant age-correlated CpG sites located at *FHL2* (6 CpGs), *EDARADD* (1 CpG) and *PDE4C* (3 CpGs). A final dual-locus model constructed with the *FHL2* CpG2 and *PDE4C* CpG3, in 21 bone samples (one sample did not amplify for *PDE4C* gene), was chosen, showing an age correlation of $R = 0.851$, highly significant (P -value = 0.000009), and explaining 69% of the variation in age (corrected $R^2 = 0.694$) (**Table 4.39**). The developed formula obtained with the regression coefficients to estimate age was the following: $(-40.785) + 109.005 \times \text{DNAm level } PDE4C \text{ CpG3} + 79.494 \times \text{DNAm level } FHL2 \text{ CpG2}$ (**Table 4.40**). The model showed a strong correlation between predicted and chronological ages (Spearman correlation coefficient, $r = 0.831$), with a MAD from chronological age of 4.67 years (RMSE = 5.98) (**Figure 4.31**) and a rate of correct predictions of 85.7%.

Table 4.39: Simple and multiple linear regression statistics of the best age predictors in *FHL2*, *EDARADD* and *PDE4C* genes using bisulfite sequencing to test for association between the DNAm levels and chronological age in bones from BDS.

Locus	CpG	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>PDE4C</i>	CpG3	Chr19:18233131	21	0.771	0.594	0.572	7.63	0.000044	5.53
<i>FHL2</i>	CpG3	Chr2:105399291	22	0.655	0.429	0.400	9.41	0.000945	7.12
<i>EDARADD</i>	CpG3	Chr1:236394382	22	-0.430	0.185	0.144	11.25	0.045937	8.57
<i>Multiple linear regression</i>									
APM (<i>FHL2</i> CpG2 and <i>PDE4C</i> CpG3)			21	0.851	0.725	0.694	6.46	0.000009	4.67

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Table 4.40: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *FHL2* and *PDE4C* selected by stepwise regression approach, in bones from BDS.

Marker	Coefficient	P-value
(Intercept)	-40.785	0.029
<i>FHL2</i> CpG2	79.494	0.009
<i>PDE4C</i> CpG3	109.005	0.000

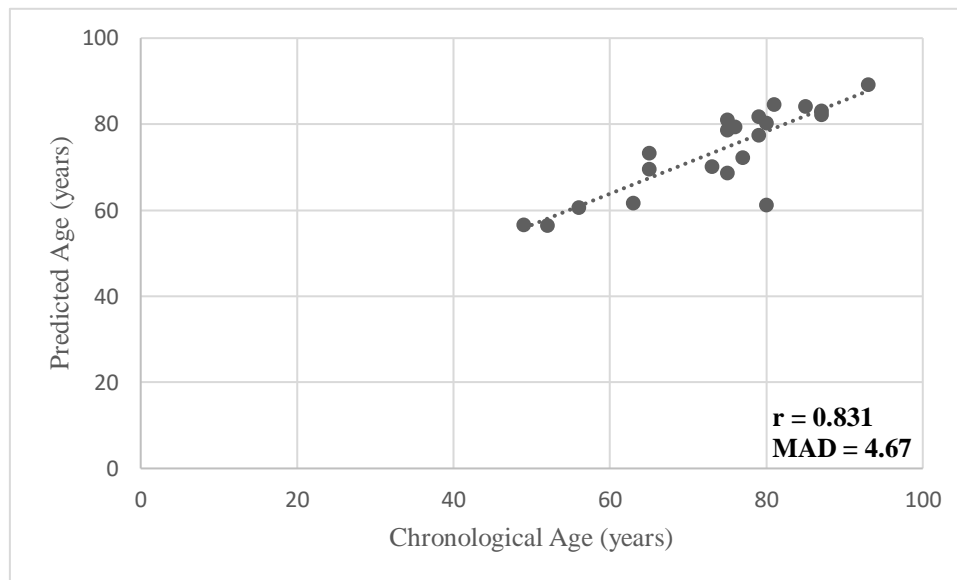


Figure 4.31: Predicted age *versus* chronological age using the two markers *FHL2* CpG2 and *PDE4C* CpG3 in bones from BDS. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Validation of the multi-locus APM developed in deceased individuals

The 3-fold cross validation allowed to estimate an averaged MAD of 6.39 years (RMSE = 6.42). The validation by splitting the training set (21 bones) into two sets of 11 and 10 samples (training and validation sets) revealing an independent MAD value for the training set of 5.28 years (RMSE = 7.20). The multiple linear regression equation developed in the training set of 10 samples was applied to the validation set, allowing to obtain a MAD of 4.90 years (RMSE = 5.17).

3.2.2. DNAm data obtained in bones from BDS using SNaPshot methodology

Using the multiplex SNaPshot assay of Jung *et al.* (2019), DNAm levels from the five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*, *TRIM59* genes were evaluated in 22 bones from BDS (10 males, 12 females; aged 49-93 years old).

Correlation between DNAm levels and chronological age

A negligible and non-significant age correlation was observed for all genes, (Table 4.41). One sample did not amplify for *ELOVL2* gene.

C. DNA methylation age estimation in fresh bone samples

Table 4.41: Simple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes using SNaPshot in bones from BDS.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>ELOVL2</i>	Chr6:11044628	21	0.412	0.170	0.124	11.75	0.071018	-
<i>FHL2</i>	Chr2:105399282	22	0.245	0.060	0.010	12.38	0.285069	-
<i>KLF14</i>	Chr7:130734355	22	0.256	0.066	0.016	12.34	0.262466	-
<i>TRIM59</i>	Chr3:160450189	22	0.147	0.022	-0.030	12.63	0.523623	-
<i>C1orf132</i>	Chr1:207823681	22	-0.040	0.002	-0.051	12.76	0.863357	-

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

3.2.3. Comparison between methodologies

For the training set of bones from BDS, comparing the correlation with age for CpGs located at *ELOVL2* (Chr6:11044628), *FHL2* (Chr2:105399282) and *C1orf132* (Chr1:207823681) genes captured by Sanger sequencing and SNaPshot methodologies we observed a low age correlation values in both methodologies, excepting for *FHL2* through Sanger sequencing showing a moderate age correlation (**Table 4.42**).

Table 4.42: Comparison of age-correlated values obtained in bones from BDS through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38 (Position in 450K array)	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Bone deceased (49-93 years)		Bone deceased (49-93 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	-0.223	0.050	0.412	0.170
<i>C1orf132</i> Chr1:207823681	-0.305	0.093	-0.040	0.002
<i>FHL2</i> Chr2:105399282 (cg06639320)	0.580	0.337	0.245	0.060

Abbreviations: R, Pearson correlation coefficient.

3.2.4. Evaluation of the effect of PMI (postmortem interval) in BDS

We selected 12 individuals from the BDS and made some pairs (A, B, C, D, E and F) with identical or similar chronological age to look for methylation differences due to the PMI (**Table 4.43**). These individuals suffer the same *postmortem* conditions of body preservation: pre-treatment of the body with Thiel Embalming (Eisma *et al.*, 2013).

Table 4.43: *Postmortem* interval of the selected BDS with the same or similar chronological age.

ID	Age	Sex	PMI*	Pair
BDS 2	65	F	3	A1
BDS 3	65	F	2	A2
BDS 5	75	F	3	B1
BDS 19	75	F	3	B2
BDS 20	79	M	2	C1
BDS 26	79	M	3	C2
BDS 21	87	F	2	D1
BDS 4	87	F	4	D2
BDS 28	80	M	3	E1
BDS 37	80	F	3	E2
BDS 30	76	M	2	F1
BDS 17	77	F	3	F2

* PMI is the time between the date at death of the individual and the date of the analysis of the sample (Annex II).

We calculated the mean of the differences between DNAm levels obtained for all the CpGs within each gene, *ELOVL2* (9 CpGs), *EDARADD* (4 CpGs), *FHL2* (12 CpGs), *PDE4C* (12 CpGs) and *C1orf132* (6 CpGs), for the individuals of each pair A, B, C, D, E and F. The obtained values were between 0.0128 and 0.1817 (**Table 4.44**).

Table 4.44: Average of the differences between DNAm levels of all the CpGs from each gene for the two individuals included in each pair A, B, C, D, E and F.

Pair	<i>ELOVL2</i>	<i>FHL2</i>	<i>EDARADD</i>	<i>PDE4C</i>	<i>C1orf132</i>
B1 + B2	0.120949	0.030042	0.048724	0.112839	0.099582
E1 + E2	0.056997	0.037493	0.023679	0.114643	0.159157
A1 + A2	0.154689	0.032505	0.057994	0.078198	0.012791*
C1 + C2	0.048475	0.08736	0.135608	0.046099	0.037883
F1 + F2	0.145283	0.034358	0.02165	0.097257	0.103971
D1 + D2	0.099053	0.038798	0.024982	0.084667	0.181683*

* The lowest and highest difference values are in bold.

4. Discussion

Anthropological age estimation approaches have been continually used in the past decades, helping forensic investigations. Forensic anthropology has been accepted as a “means of personal identification” in cases in which the traditional identifiers, as DNA analysis, cannot be applied (de Boer *et al.*, 2020). In recent years, the growth of DNAm age has brought new knowledge and challenges in the forensic field (epigenetic approaches). The analysis of DNAm levels in different tissue types allowed the development of several epigenetic clocks (review in Goel *et al.*, 2017; Zolotareno *et al.*, 2019). Despite this, only few studies have considered the study of DNAm levels in bone samples (Horvath *et al.*, 2015, 2018; Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020).

In this study, DNAm levels of 43 CpGs located at the five high age-correlated genes (*ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132*) were evaluated using Sanger sequencing methodology in two training sets of bone samples (31 bones from autopsies and 22 bones from BDS). Using the stepwise regression approach with methylation information of 29 significant CpGs located at the *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes analyzed through the Sanger sequencing methodology, a multi-locus APM with six markers (*ELOVL2* CpG5, *ELOVL2* CpG6, *ELOVL2* CpG7, *C1orf132* CpG1, *EDARADD* CpG3 and *EDARADD* CpG4) was developed in the training set of 29 bones from autopsies. This final APM revealed a high age correlation value ($R = 0.970$), showing high accuracy allowing to obtain a MAD from chronological age of 2.56 years. This suggests that bisulfite Sanger sequencing methodology could be suited for forensic purposes using bone derived DNAm. In the training set of bones from BDS, through the Sanger sequencing methodology, the stepwise regression approach using the 10

significant age-correlated CpGs located at *FHL2*, *EDARADD* and *PDE4C* genes, allowed to obtain a dual-locus APM with *FHL2* CpG2 and *PDE4C* CpG3. The model showed a high age-correlated value ($R = 0.851$), allowing to obtain a MAD from chronological age of 4.67 years.

Using the SNaPshot methodology, five CpGs located at *ELOVL2*, *C1orf132*, *FHL2*, *KLF14* and *TRIM59* genes were analyzed in these same training sets. The SNaPshot method allowed to develop in bones from autopsies a final dual-locus APM, combining *FHL2* and *KLF14* genes, showing moderate age correlation value of 0.777 and a MAD from chronological age of 7.18 years. As this model included only two CpG sites, while the APM constructed with methylation information captured by Sanger sequencing used six predictor variables, this should justify the higher MAD value obtained in the APM developed through the SNaPshot assay.

Despite the scarcity of methylation data reported in bones samples, Naue *et al.* (2018) investigated through massive parallel sequencing whether 13 previously selected age-dependent loci have predictive value in several forensically relevant tissues including bones. Using 29 bones from deceased individuals (aged 0–87 years old) they found no age-dependency for several genes, but statistically significant age correlations of 0.58, 0.51 and 0.61 (Pearson correlation) were observed for one CpG site within the amplified region of genes *ELOVL2*, *KLF14* and *TRIM59*. This is in concordance with our study, in which methylation levels captured by the SNaPshot assay in bone samples from autopsies, revealed for CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes significant correlation values ($R \geq 0.540$). Moreover, one CpG located at *KLF14* was included in our final dual-locus APM.

Gopalan *et al.* (2019) generated DNAm data from 32 bones alongside with published data from 133 bones from living and deceased individuals (aged 49-112 years

old). They identify 108 sites with significant association with age. The authors have developed an APM that produces high accuracy of age estimation. From all CpG sites that comprised the best model (37 bone clock CpG regions), two *ELOVL2* positions (cg16867657 and cg24724428) and one *KLF14* position (cg07955995) were included. In particular, the *ELOVL2* CpG6 (Chr6:11044644; cg16867657) from our APM with six CpGs developed by Sanger sequencing has already been included in Gopalan *et al.* (2019) and Naue *et al.* (2018) studies, revealing to be a promising marker for bone samples. These results support the notion that gene *ELOVL2* is a highly age-correlated gene in many tissues (Gopalan *et al.*, 2019).

Recently, Lee and collaborators (2020) applied the previously developed “skin & blood clock” focused on 391 CpGs proposed by Horvath *et al.* (2018) to 12 bone samples. A model accuracy of 6.4 years ($r = 0.964$) was obtained. This value is similar to the MAD value obtained in the dual-locus model (*FHL2* and *KLF14*) developed in the training set of bones from autopsies through the SNaPshot methodology (MAD of 7.18 years, $R = 0.777$). In addition, Lee and collaborators (2020) observed a strong age correlation value for several CpGs including cg06639320 (*FHL2*, Chr2:105399282, $R = 0.810$), cg09809672 (*EDARADD*, Chr1:236394382, $R = -0.743$), cg21572722 (*ELOVL2*, Chr6:11044661, $R = 0.837$), cg16867657 (*ELOVL2*, Chr6:11044644, $R = 0.910$), cg20426994 (*KLF14*, Chr7:130733449, $R = 0.881$) and cg07553761 (*TRIM59*, Chr3:160450189, $R = 0.702$). Interestingly, in our study for the set of bones from autopsies, comparable age-correlated values were obtained for these same positions using the Sanger sequencing, *FHL2* CpG1 (Chr2:105399282, $R = 0.692$), *EDARADD* CpG3 (Chr1:236394382, $R = -0.561$), *ELOVL2* CpG9 (Chr6:11044661, $R = 0.590$) and *ELOVL2* CpG6 (Chr6:11044644, $R = 0.852$), or using the SNaPshot methodology, *FHL2* (Chr2:105399282, $R = 0.708$) and *TRIM59* (Chr3:160450189, $R = 0.633$).

In the training set of bones from BDS no association was observed between DNAm levels and sex, in all the investigated CpGs for both methodologies, as previously reported for these genes or others, and other tissue types (Bekaert *et al.*, 2015a; Huang *et al.*, 2015; Freire-Aradas *et al.*, 2018; Daunay *et al.*, 2019; Márquez-Ruiz *et al.*, 2020). DNAm sex analysis was not performed in set of bones from autopsies due the limited number of females. Additionally, as both training sets of bones were not representative of all age groups, namely groups aged 26-81 years old and 49-93 years old, we did not analyze the differences between predicted and chronological ages with the increase of age.

Comparing both methodologies of bisulfite sequencing and SNaPshot, in the set of bones collected during autopsies, similar age correlation values were observed for *ELOVL2* and *FHL2* genes. However, *C1orf132* gene revealed a stronger age correlation value in Sanger sequencing. For the training set of bones from BDS a weak or negligible age correlation was obtained for the same position from *ELOVL2* and *C1orf132* genes in both methodologies. For the CpG located at *FHL2* a stronger age correlation value was observed using Sanger sequencing methodology. As we use a limited set of bone samples in both methodologies, these differences in DNAm levels can be related with the number of samples. In future studies, the use of a larger sample set will allow to explore the differences between methodologies more accurately.

Significant differences in DNAm levels were observed between the two sets of bones, from BDS and from autopsies, which led us to make an independent analysis for each set of bone samples and consequently, to the development of different APMs for each group. We hypothesize that these differences between the two sets of bones can be related with the treatment of the BDS after death. As BDS need to be preserved for long periods after death, the body is submitted to an embalming process: Thiel Embalming

(Eisma *et al.*, 2013). This treatment included chemical components, which can influence DNAm levels. Although this topic has not been investigated so far, it is a relevant question for forensic investigations.

In addition, we investigated if the different PMI of the individuals could interfere in DNAm levels. Bones from autopsies were collected until five days after death and bones from BDS have a higher PMI. We cannot compare samples from BDS with samples from autopsies because these samples do not share the same preservation conditions (pre-treatment of the body), and this can also interfere in DNAm levels as previously described. For assessment of PMI effect, we compared pairs of bones from BDS with the same preservation conditions and the same or similar chronological age (consequently with presumed similar DNAm levels), being the PMI the only factor of DNAm changes. As the difference obtained in DNAm levels is very low, we can hypothesize that there is no effect of PMI in DNAm levels in bones samples from BDS. We can observe that the mean difference of DNAm levels for pairs B and E (which have the same PMI between the two individuals of each pair) is similar to the obtained for the individuals of pair D, which has the higher difference on PMI (2 years).

In conclusion, considering that to date only few reports used bone samples in development of models for forensic age estimation, we evaluated DNAm levels from two sets of bones using the bisulfite Sanger sequencing and SNaPshot methodologies. Our study allowed to develop a highly accurate APM in bone samples from autopsies with six CpGs located at genes *ELOVL2*, *EDARADD* and *C1orf132* through bisulfite Sanger sequencing, with a MAD from chronological age of 2.56 years. Moreover, in a training set of bones from BDS, a dual-locus model with *FHL2* and *PDE4C* genes with an accuracy of 4.67 years was developed using bisulfite sequencing. The SNaPshot method allowed to construct a final dual-locus model with *FHL2* and *KLF14* genes for bones

from autopsies, exhibiting moderate age prediction accuracy (MAD = 7.18 years). Evaluating DNAm changes by SNaPshot assay in the training set of bones from BDS any APM could be developed. We observed differences in DNAm levels of bones from autopsies and bones from BDS, which can be related with the treatment of the body after death. Meanwhile, this issue needs to be explored in future studies. Our study suggests that skeletal human remains, with high resistance to harsh conditions and often recoverable for long *postmortem* intervals, can be prime targets for DNAm analyses in forensic contexts.

Chapter 4. Results and discussion

D. DNA methylation age estimation in buccal swabs

1. Introduction

Nowadays, age estimation in the living is increasingly important in forensic contexts. Recently, DNA methylation (DNAm) age estimation appears as promising approach for several forensic contexts. Most epigenetic age clocks have been developed on different tissue types from living individuals, such as blood, teeth, buccal swabs and saliva, using many age-correlated markers (as *ELOVL2*, *FHL2*, *EDARADD*, *ASPA*, *PDE4C*, *PENK*, *C1orf132*, *TRIM59* and *KLF14* genes) and different technologies including SNaPshot and pyrosequencing (e.g. Garagnani *et al.*, 2012; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Giuliani *et al.*, 2016; Park *et al.*, 2016; Eipel *et al.*, 2016; Jung *et al.*, 2019; Pfeifer *et al.*, 2020; Márquez-Ruiz *et al.*, 2021).

The collection of buccal swabs and saliva samples does not require invasive procedures so it can be a good approach for forensic DNAm age estimation. In fact, several epigenetic clocks have been developed using buccal swabs and saliva (Bocklandt *et al.*, 2011; Park *et al.*, 2014; Silva *et al.*, 2015; Bekaert *et al.*, 2015b; Eipel *et al.*, 2016; Hong *et al.*, 2017; Alghanim *et al.*, 2017; Hamano *et al.*, 2017; Jung *et al.*, 2019; Koop *et al.*, 2020; Pfeifer *et al.*, 2020).

Bekaert *et al.* (2015b) using pyrosequencing evaluated DNAm levels of *ASPA*, *ELOVL2*, *PDE4C* and *EDARADD* genes in 50 buccal swabs from living individuals obtained a Mean Absolute Deviation (MAD) between chronological and predicted ages of 3.32 years.

Eipel *et al.* (2016), using pyrosequencing, tested in 55 buccal swabs from healthy individuals (1-85 years) the applicability of Weidner's model developed for blood samples with three CpGs in *PDE4C*, *ASPA* and *ITGA2B* genes, which revealed a MAD from chronological age of 5.0 years ($R^2 = 0.81$) (Weidner *et al.*, 2014). Using the age

prediction formula from the Weidner's model, a lower accuracy was obtained by Eipel *et al.* (2016) for buccal swabs (MAD = 14.6 years, $R^2 = 0.91$). Therefore, Eipel and collaborators (2016) developed a specific age prediction model (APM) with the previously reported three CpG sites for the 55 buccal swabs obtaining a higher MAD value of 4.3 years.

Using the SNaPshot method, Jung *et al.* (2019) developed tissue-specific APMs for blood, saliva and buccal swabs, obtaining more similar MAD values for blood and saliva samples (MAD = 3.17 years; MAD = 3.29 years, respectively) than the obtained in buccal swabs (MAD = 3.82 years). Moreover, the age correlation values obtained for saliva and blood samples in CpGs located at *ELOVL2* (Chr6:11044628), *FHL2* (Chr2:105399282), *Clorf132* (Chr1:207823681) genes were more similar ($0.699 \leq R \leq 0.832$ and $0.637 \leq R \leq 0.893$, respectively) than the obtained for buccal swabs ($0.314 \leq R \leq 0.684$).

Naue *et al.* (2018) investigated DNAm levels of 13 age-correlated markers in blood, bone, brain, buccal swabs and muscle from 29 deceased individuals. For buccal swabs, strong and significant age correlation values were obtained for CpGs located at *ELOVL2* (Chr6:11044644; cg16867657), *KLF14* (Chr7:130734357) and *TRIM59* (Chr3:160450189; cg07553761) ($R = 0.83$; $R = 0.70$; $R = 0.85$, respectively).

In 2020, Pfeifer and colleagues investigated DNAm levels of *ELOVL2*, *EDARADD*, *PDE4C* and *ASPA* genes by pyrosequencing in 149 buccal swabs from living individuals. Using 100 buccal swabs and applying the published equation developed by Bekaert *et al.* (2015b), a MAD of 8.68 years was obtained. Pfeifer *et al.* (2020) developed a new specific model for their training set of buccal swabs with some CpGs from the same genes investigated by Bekaert *et al.* (2015b), obtaining a better accuracy with a MAD of 4.65 years.

More recently, a study proposed by Koop *et al.* (2020) evaluated DNAm changes of one CpG from *PDE4C* gene located upstream of cg17861230 in buccal swabs from living and deceased individuals using pyrosequencing. They observed a higher age correlation value in the model developed for the training set of 71 living individuals ($r^2 = 0.87$) with only this site from *PDE4C*. Testing this APM to buccal swabs from an independent sample set of living individuals ($N = 71$, $r^2 = 0.85$) and from deceased individuals ($N = 52$, $r^2 = 0.90$), a moderate accuracy was observed (MAD = 7.8 years and MAD = 9.1 years, respectively). In the set of deceased individuals, Koop and colleagues considerate different stages of body decomposition to evaluate *postmortem* DNAm changes, observing no influence in age estimation.

Considering the potential usefulness of buccal swabs as a source of DNA for DNAm age clocks, this study aimed to examine in a sample set of buccal swabs from living Portuguese individuals: i) the DNAm profiles for age prediction purposes of CpG sites from *ELOVL2* locus, one of the most powerful age predictor markers proposed in previous studies, through the bisulfite PCR sequencing methodology; and ii) the DNAm levels of three CpGs located at *TRIM59*, *KLF14* and *ELOVL2* genes, using the multiplex SNaPshot method proposed by Jung *et al.* (2019). These three CpGs were the best age-correlated markers addressed in buccal swabs by Jung *et al.* (2019).

2. Materials and Methods

2.1. Sample collection

A set of 39 buccal swabs from healthy individuals of Portuguese ancestry (16 males, 23 females; aged 3-86 years old) was collected according to recommended

guidelines. Written informed consent was previously obtained from adult participants and from children's parents under the age of 18 years.

2.2. DNA extraction, quantification and bisulfite conversion

Buccal swabs collected with a sterile brush (Sarstedt, Nümbrecht, Germany) were submitted to DNA extraction using the *FavorPrepTM Genomic DNA mini kit* (Favorgen Biotech Corp, Taiwan) and DNA extracts were quantified in a Nanodrop spectrophotometer (Thermo Fisher Scientific), according to previously described in *Chapter 3. Sample and design research*.

Genomic DNA was subjected to bisulfite conversion using *EZ DNA Methylation-GoldTM Kit* (Zymo Research, Irvine, USA), as previously described in *Chapter 3. Sample and design research*. Briefly, 20 µl of genomic DNA (in a total amount of 40 to 400 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 µl.

2.3. Polymerase chain reaction (PCR) and Sanger sequencing

After bisulfite conversion, the modified DNA samples were submitted to PCR for nine CpGs from *ELOVL2* gene using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described in *Chapter 3. Sample and design research*.

2.4. SNaPshot assay

After bisulfite conversion, the modified DNA samples were submitted to a multiplex SNaPshot assay for three CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes

with the primers and conditions previously described in Jung *et al.* (2019). Particular conditions for multiplex PCR amplification and multiplex SBE reactions were described in *Chapter 3. Sample and design research.*

2.5. DNAm quantification

DNAm quantification in both methodologies was assessed as previously described in *Chapter 3. Sample and design research.*

2.6. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Independent analyses were made for the data obtained through Sanger sequencing and SNaPshot methodologies. Linear regression models were used to analyze relationships between methylation levels and chronological age as previously described in *Chapter 3. Sample and design research.* Briefly, using the simple linear regression coefficients from each significant age-correlated CpG site, we predicted age of individuals for each individual gene. The best combination of significant age-correlated CpGs selected from a stepwise regression analysis was used in a multiple regression approach to build the final multi-locus APM in each set of buccal swabs.

The MAD between chronological and predicted ages and the root mean square error (RMSE) were calculated for the overall training set using the final APM developed in each methodology. Moreover, MAD values were calculated for subsets of three distinct age categories in the training set (<13 years old; 14-33 years old; 34-86 years old). Each obtained result was interpreted as either correct or incorrect if the predicted age was

concordant with the chronological age using a cutoff value according to the standard error (SE) of estimate obtained in the developed APM.

Validation of the final APMs developed for each set of buccal swabs in both methodologies was performed by splitting the complete data set into two subsets (training and validation sets) and by 3-fold cross validation, as previously described in *Chapter 3. Sample and design research.*

The evaluation of differences between sex was made through comparison of two regression lines relating chronological age and DNAm levels of each gene at two levels (males/females) of the categorical factor, using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA) for both methodologies.

3. Results

DNAm levels of nine CpGs located at the *ELOVL2* locus, analyzed through bisulfite PCR sequencing, and DNAm levels of three CpGs located at the *ELOVL2*, *KLF14* and *TRIM59* genes, examined by the multiplex methylation SNaPshot assay described by Jung *et al.* (2019), were assessed in buccal swabs from Portuguese volunteers.

The accuracy of measurement of DNAm levels obtained by Sanger sequencing methodology was tested using DNAm standards of 0%, 50% and 100%. The obtained DNAm levels of the best-selected *ELOVL2* sites in buccal swabs showed a significant linear relationship to the expected DNAm levels (**Supplementary Figure S27**).

3.1. DNAm data obtained in buccal swabs using bisulfite Sanger sequencing

Correlation between DNAm levels and chronological age

Testing the association between DNAm levels of CpGs located at the *ELOVL2* locus and chronological age in a training set of 23 buccal swabs from living Portuguese individuals (15 females, 8 males; aged 3-86 years old), positive age correlations were observed for all the CpGs (**Supplementary Table S15**). The best CpGs at *ELOVL2* are presented in **Supplementary Figure S28**.

No statistically significant difference was observed comparing DNAm levels between males and females for the target sites selected in *ELOVL2* locus (P -value >0.05 , **Table 4.45**). Therefore, all the subsequent analyses were made ignoring differences between males and females.

Table 4.45: Comparison of two regression lines between males and females in buccal swabs from living individuals using data obtained from Sanger sequencing.

<i>ELOVL2</i>	<i>P</i> -value	
	Intercept	Slope
CpG1	0.5697	0.9452
CpG4	0.5400	0.1776
CpG5	0.4726	0.7819

Development of an age prediction model (APM)

Simple linear regression models testing the correlation between DNAm levels and chronological age, revealed significant associations for all the CpGs located at *ELOVL2* ($R \geq 0.485$) (**Supplementary Table S15**). The strongest age correlation value was observed for *ELOVL2* CpG1 ($R = 0.823$, P -value = 0.000001), explaining 66.2% of the variation in age, followed by *ELOVL2* CpG5 ($R = 0.806$, P -value = 0.000003), explaining 63.3% of the variation in age, and *ELOVL2* CpG4 ($R = 0.787$, P -value = 0.000008),

explaining 60.1% of the variation in age (**Table 4.46; Supplementary Figure S28**). The remaining CpG sites showed moderate age-correlated values ($0.485 \leq R \leq 0.683$) (**Supplementary Table S15**).

Page | 230

Using the simple linear regression coefficients for the three strongest *ELOVL2* age-associated markers, the predicted age of the individuals was calculated and the obtained MAD values were: 10.29 years for *ELOVL2* CpG1; 10.62 years for *ELOVL2* CpG5; and 11.39 years for *ELOVL2* CpG4 (**Supplementary Figure S29**).

Multiple linear regression analysis with simultaneously methylation information of CpG1, CpG4 and CpG5 located at *ELOVL2* gene revealed a non-significant p-value for *ELOVL2* CpG5 (P -value >0.05 , **Supplementary Table S16**). Moreover, in stepwise regression analysis using simultaneously all the nine CpGs located at *ELOVL2* gene, only the *ELOVL2* CpG1 and *ELOVL2* CpG4 were selected. A multiple linear regression model with these two CpGs revealed a strong age correlation value ($R = 0.894$), explaining 78% of the variation in age (corrected $R^2 = 0.780$), highly significant (P -value = 1.0347×10^{-7}) (**Table 4.46**).

Table 4.46: Simple and multiple linear regression statistics of the best age predictors in *ELOVL2* gene to test for association between the DNAm levels obtained by bisulfite sequencing and chronological age in buccal swabs from living individuals.

<i>ELOVL2</i>	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
CpG1	Chr6:11044628	23	0.823	0.677	0.662	14.40	0.000001	10.29
CpG4	Chr6:11044640	23	0.787	0.619	0.601	15.65	0.000008	11.39
CpG5	Chr6:11044642	23	0.806	0.650	0.633	14.99	0.000003	10.62
<i>Multiple linear regression</i>								
APM (<i>ELOVL2</i> CpG1 and <i>ELOVL2</i> CpG4)		23	0.894	0.800	0.780	11.62	1.035×10^{-7}	8.32

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

D. DNA methylation age estimation in buccal swabs

Using the multiple linear regression coefficients (**Table 4.47**) the developed equation to estimate age was the following: $(-88.854) + 92.953 \times \text{DNAm level } ELOVL2 \text{ CpG1} + 81.656 \times \text{DNAm level } ELOVL2 \text{ CpG4}$. Predicted and chronological ages of the 23 buccal swabs from Portuguese individuals were highly correlated (Spearman correlation coefficient, $r = 0.917$) (**Figure 4.32**). The difference between predicted and chronological ages for each individual using the two-CpGs APM developed from *ELOVL2* allowed to obtain a MAD of 8.32 years (RMSE = 10.07) (**Table 4.46**). The percentage of correct predictions (considering cutoff of 11.62 years) was 73.9%.

Table 4.47: Statistical parameters obtained in a multiple regression model with the two CpGs in *ELOVL2* gene, selected by stepwise regression approach, in buccal swabs from living individuals.

<i>ELOVL2</i>	Coefficient	P-value
(Intercept)	-88.854	0.000
CpG1	92.953	0.000
CpG4	81.656	0.002

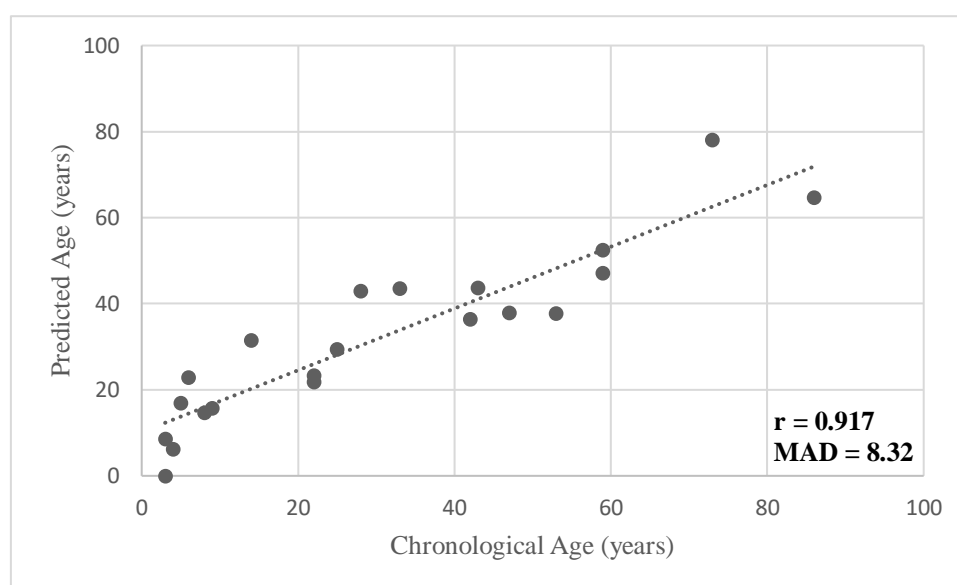


Figure 4.32: Predicted age *versus* chronological age using the two markers *ELOVL2* CpG1 and *ELOVL2* CpG4 in buccal swabs from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological ages with aging

Some differences between predicted and chronological ages can be observed with the increase of age using the multiple APM developed with CpGs from *ELOVL2* (**Figure 4.33**).

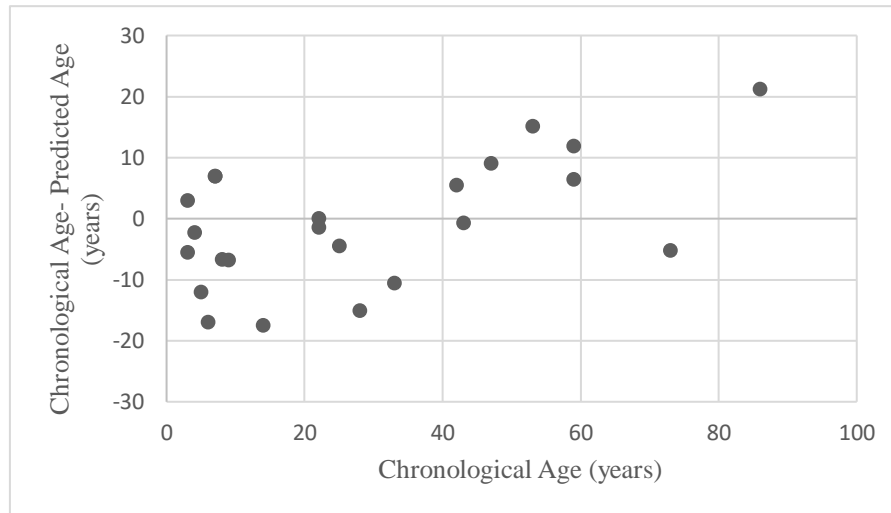


Figure 4.33: Differences between chronological and predicted ages (years) plotted against chronological age (years) in buccal swabs from living individuals.

If three age groups (Group 1: <13 years old; Group 2: 14-33 years old; Group 3: 34-86 years old) were considered from the overall training set to evaluate age-related changes, and the MAD values and percentage of correct predictions were calculated within each group, we can observe an increase in the MAD values from G1 to G3 (**Table 4.48; Figure 4.34**).

Table 4.48: MAD between predicted and chronological ages stratified by age group in the training set of 23 buccal swabs from living individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<13 years	9	7.45	77.8
G2	14-33 years	6	8.15	66.6
G3	34-86 years	8	9.42	62.5
Overall	3-86 years	23	8.32	73.9

In concordance, the lower percentage of correct predictions was observed in the older age G3 (62.5%) and the higher values of correct predictions were observed in the younger age categories: G1 (77.8%) and G2 (66.6%) (**Table 4.48**).

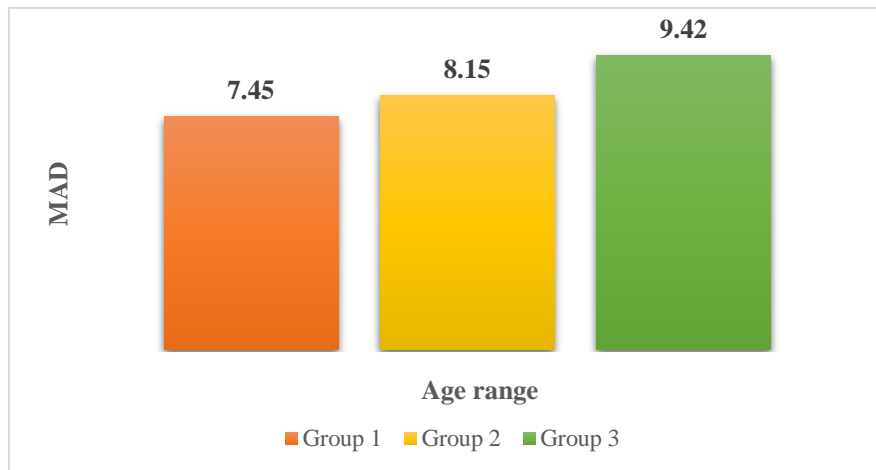


Figure 4.34: MAD from chronological age calculated for each age group in buccal swabs from living individuals. MAD increases with age. The MAD is printed on top of each respective age range.

Validation of the multiple APM developed in living individuals

Through the 3-fold cross validation, each independent multiple linear regression predictive model tested in the removed samples (three validation sets) allowed to obtain an averaged MAD of 10.52 years (RMSE = 10.56). The validation by splitting the sample into two sets of 12 and 11 samples (training and validation sets) allowed to obtain a MAD value of 8.94 years (RMSE = 10.52) for the training set, and a MAD of 10.45 years (RMSE = 12.38) for the validation set.

3.2. DNAm data obtained in buccal swabs using SNaPshot methodology

Correlation between DNAm levels and chronological age

DNAm levels of CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes were analyzed using the multiplex SNaPshot assay in 39 buccal swabs from living individuals (16 males, 23 females; aged 3-86 years old). Some samples have unsuccessful SNaPshot sequencing: four samples for *TRIM59*, four other samples for *KLF14*, one of these four samples did not amplify for *ELOVL2* and two other samples did not amplify for *ELOVL2*.

Evaluating DNAm association with age, positive correlations were observed for the CpGs located at all genes (**Supplementary Figure S30**).

DNAm levels and sex

Using DNAm information of CpGs located at *ELOVL2*, *KLF14* and *TRIM59* no statistically significant differences were observed between males and females in the training set (P -value >0.05 , **Table 4.49**).

Table 4.49: Comparison of two regression lines between males and females in buccal swabs from living individuals using data obtained from the SNaPshot methodology.

Marker	P-value	
	Intercept	Slope
<i>ELOVL2</i>	0.9192	0.0751
<i>TRIM59</i>	0.9063	0.1896
<i>KLF14</i>	0.8396	0.3157

Development of an age prediction model (APM)

Testing association between DNAm levels and chronological age, the strongest correlation was observed for *TRIM59* marker ($R = 0.946$, P -value = 1.183×10^{-17}), explaining 89.1% of the variation in age, followed by *ELOVL2* ($R = 0.846$, P -value = 8.295×10^{-11}), explaining 70.7% of the variation in age, and *KLF14* ($R = 0.821$, P -value

D. DNA methylation age estimation in buccal swabs

$= 9.204 \times 10^{-10}$), explaining 67.4% of the variation in age. The predicted age of individuals was calculated through the simple linear regression coefficients for the individual markers allowing to obtain a MAD of 6.73 years for *TRIM59*, 11.44 years for *ELOVL2* and 11.86 years for *KLF14* (**Table 4.50; Figure 4.35; Supplementary Figure S31**).

Table 4.50: Simple linear regression statistics at the three CpGs of the *ELOVL2*, *KLF14* and *TRIM59* genes using SNaPshot assay in buccal swabs from living individuals.

Locus	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
<i>TRIM59</i>	Chr3:160450189	35	0.946	0.894	0.891	8.98	1.183×10^{-17}	6.73
<i>ELOVL2</i>	Chr6:11044628	36	0.846	0.716	0.707	15.27	8.295×10^{-11}	11.44
<i>KLF14</i>	Chr7:130734355	35	0.821	0.684	0.674	15.44	9.204×10^{-10}	11.86

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

Using simultaneously these three CpGs through multiple linear regression statistics, only the CpG located at *TRIM59* gene revealed a significant age correlation value (P -value = 3.074×10^{-8}) (**Supplementary Table S17**). In concordance, by stepwise regression approach, only this same CpG was chosen. Age prediction estimated according to the individual regression coefficients for the highest age-associated marker, *TRIM59* (**Table 4.51**) was as follows: $(-8.920) + 108.282 \times \text{DNAm level } TRIM59$. The obtained Spearman correlation value between predicted and chronological ages was 0.962 and the MAD from chronological age was 6.73 years (RMSE = 8.70) (**Table 4.50; Figure 4.35**). The percentage of correct predictions (considering the cutoff of 8.98) was 80%.

Table 4.51: Statistical parameters obtained by simple linear regression for the *TRIM59* gene, selected by stepwise regression approach in buccal swabs.

Marker	Coefficient	<i>P</i> -value
(Intercept)	- 8.920	0.005
<i>TRIM59</i>	108.282	0.000

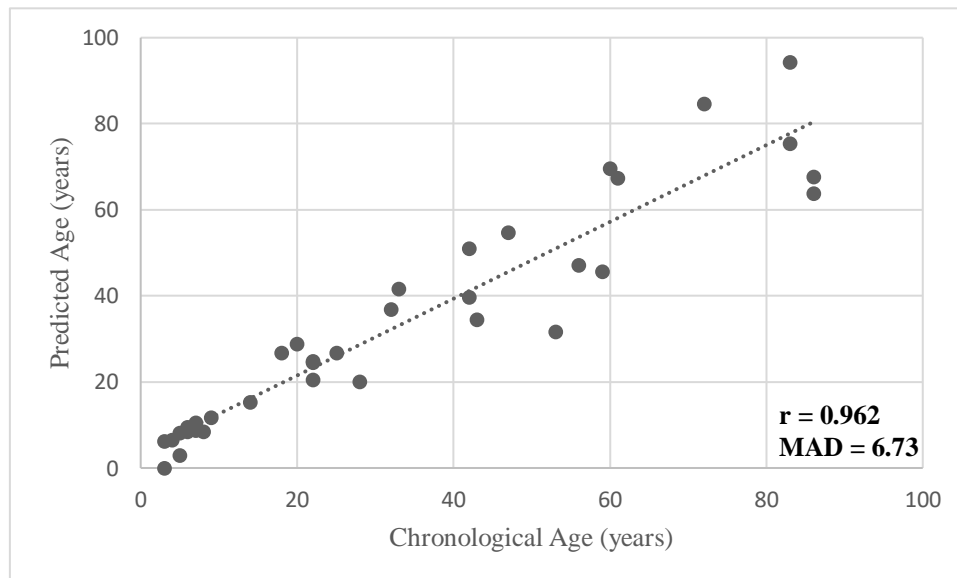


Figure 4.35: Predicted age *versus* chronological age using the multiplex methylation SNaPshot assay at the CpG site located at *TRIM59* gene in buccal swabs from living individuals. MAD and Spearman correlation coefficient, r , are plotted on the chart.

Differences between predicted and chronological age with aging

Using the developed simple linear regression model with methylation information of *TRIM59* captured by SNaPshot methodology, we observed some differences between predicted and chronological ages with increasing age (**Figure 4.36**).

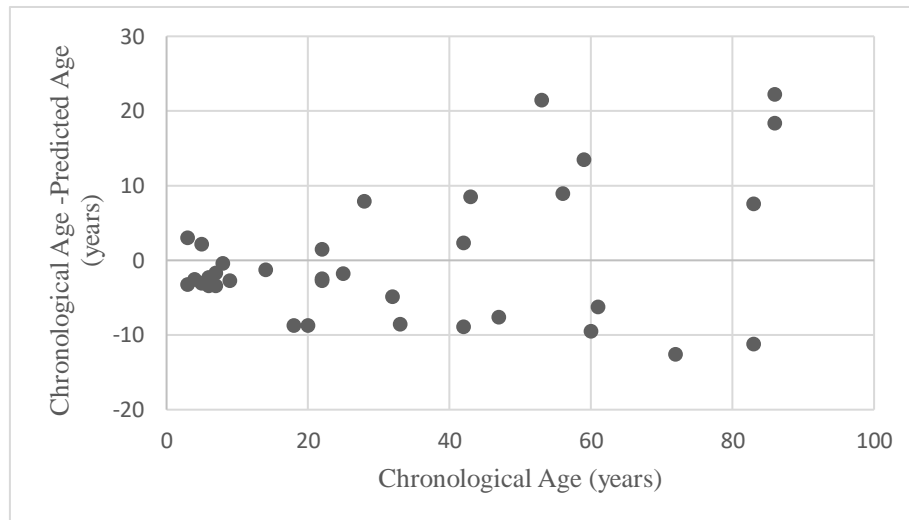


Figure 4.36: Differences between chronological and predicted ages (years) plotted against chronological age (years) in buccal swabs from living individuals.

We divided our training set of 35 buccal swabs successful amplified and sequenced by SNaPshot assay for the *TRIM59* in three age ranges (Group 1: <13 years old; Group 2: 14-33 years old; Group 3: 34-86 years old) (**Table 4.52; Figure 4.37**) and calculated the MAD and percentage of correct predictions within each group. The obtained MAD value increases from G1 to G3 with the increasing of age (MAD = 2.55 years, MAD = 5.24 years and MAD = 10.69 years, respectively). In concordance, the lower percentage of correct predictions was observed in the older group G3: 53% (**Table 4.52; Figure 4.37**).

Table 4.52: MAD between predicted and chronological ages stratified by age group in the training set of 35 buccal swabs from living individuals.

Group	Age range	N	MAD	Correct Predictions (%)
G1	<13 years	11	2.55	100
G2	14-33 years	9	5.24	100
G3	34-86 years	15	10.69	53
Overall	3-86 years	35	6.73	80

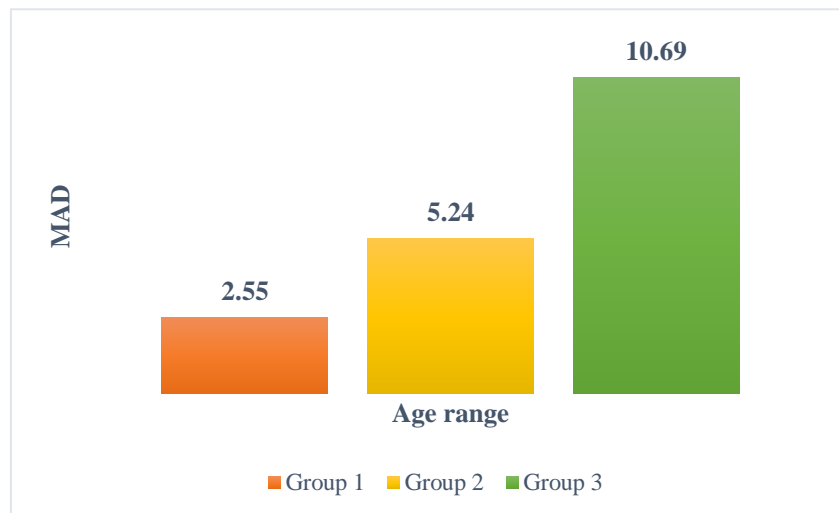


Figure 4.37: MAD from chronological age calculated for each age group in buccal swabs from living individuals. MAD increases with age. The MAD is printed on top of each respective age range.

Validation of the simple APM developed in living individuals

The 3-fold cross validation allowed to estimate an averaged MAD in the three validation sets of 7.07 years (RMSE = 7.17). The validation by splitting the sample in the training set of 19 samples and in the validation set of 16 samples allowed to obtain an independent MAD value of 5.98 years (RMSE = 7.68) for the training set, and MAD of 7.50 years (RMSE = 9.83) for the validation set. Both independent MAD values were very close to the MAD of 6.73 years (RMSE = 8.70) obtained from the whole data set.

3.3. Comparison between methodologies

Comparing the value of correlation with age of CpG located at *ELOVL2* using Sanger and SNaPshot methodologies, we observed similar values (**Table 4.53**).

Table 4.53: Comparison of age-correlated values obtained in buccal swabs through Sanger and SNaPshot methodologies.

Chromosomal location GRCh38	Portuguese ancestry			
	Sanger sequencing		SNaPshot	
	Buccal swabs (3-86 years old)		Buccal swabs (3-86 years)	
	R	R ²	R	R ²
<i>ELOVL2</i> Chr6:11044628	0.823	0.677	0.846	0.716

Abbreviations: R, Pearson correlation coefficient.

4. Discussion

Buccal swabs can be a relevant source of DNA, however few studies have developed DNAm clocks using this tissue type (Bekaert *et al.*, 2015b; Eipel *et al.*, 2016; Naue *et al.*, 2018; Jung *et al.*, 2019; Koop *et al.*, 2020). It is known that DNAm patterns of buccal epithelial and blood cells may differ and consequently, the epigenetic clocks developed for blood samples (as the model of Weidner *et al.*, 2014) need to be adapted when applied to other tissue type, as buccal swabs (Eipel *et al.*, 2016). It has been observed that APMs developed specifically for buccal swabs are more advantageous for accurate age predictions in buccal swabs than the use of models that were previously developed in blood (Eipel *et al.*, 2016). The specific cellular composition of saliva, blood and buccal swabs could influence age predictions based on tissue-specific CpGs (Eipel *et*

al., 2016; Vidaki *et al.*, 2016). For instance, saliva and buccal swabs can show different DNAm patterns as a consequence of the different amount of leukocytes and epithelial cells they contain (Thiede *et al.*, 2000; Vidaki *et al.*, 2016). Moreover, as saliva can have large number of leukocytes, DNAm levels of saliva and blood samples can be more similar, being reflected in the accuracy of developed models (Vidaki *et al.*, 2016; Jung *et al.*, 2019).

In the present study, we assessed DNAm levels of *ELOVL2*, *KLF14* and *TRIM59* genes in 39 buccal swabs from living Portuguese individuals through Sanger sequencing methodology (nine CpGs located at *ELOVL2* gene) and by multiplex SNaPshot method (three CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes).

Evaluating DNAm levels of CpGs from *ELOVL2* gene through Sanger sequencing in a training set of 23 buccal swabs, a two-CpGs APM combining *ELOVL2* CpG1 and *ELOVL2* CpG4 through multiple linear regression analysis allowed to obtain a strong age correlation value ($R = 0.894$), highly significant ($P\text{-value} = 1.035 \times 10^{-7}$), and explaining 78% of the variation in age. The model showed a moderate accuracy with a MAD from chronological age of 8.32 years.

We tested the experiments of Jung *et al.* (2019) evaluating simultaneously three CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes through SNaPshot assay in a training set of 39 buccal swabs. When using a multiple linear regression approach with the three loci, absence of significant values was observed for *ELOVL2* and *KLF14* loci. The predicted age of individuals calculated through the simple linear regression coefficients for the strongest age-correlated CpG located at the *TRIM59* gene ($R = 0.946$) allowed to obtain a MAD from chronological age of 6.73 years. Jung and colleagues (2019) investigating DNAm levels of *ELOVL2*, *KLF14*, *FHL2*, *TRIM59* and *C1orf132* genes in buccal swabs from Korean people observed the highest age-correlated value also

for *TRIM59* locus ($R = 0.734$), following by *KLF14* ($R = 0.711$), *ELOVL2* ($R = 0.684$), *FHL2* ($R = 0.455$) and *C1orf132* ($R = -0.314$). They developed a specific APM for buccal swabs with these five CpGs, revealing a high age correlation ($R = 0.943$) and a MAD from chronological age of 3.823 years. The MAD value reported by Jung *et al.* (2019) is lower than the obtained in our study (6.73 years), revealing higher accuracy in Korean people, which can be explained by the use of more CpGs/genes (five CpGs) in the model. Despite this, it is possible to observe higher age correlation values of the same positions for Portuguese training set (*TRIM59*, $R = 0.946$; *ELOVL2*, $R = 0.846$; *KLF14*, $R = 0.821$) comparing with the previously reported in Korean. This suggests that putative population specific changes can influence DNAm levels.

The previously reports by Bekaert *et al.* (2015b) and Eipel *et al.* (2016), using around 50 buccal swabs and the pyrosequencing methodology, revealed an accurate MAD values of 3.32 years (eight CpGs) and 4.3 years (three CpGs), respectively. In our study, the obtained MAD value is higher with both methodologies, but less CpGs were used in each APM (Sanger sequencing with two CpGs, MAD = 8.32 years; SNaPshot with only one CpG, MAD = 6.73 years). The use of less predictor variables and, in addition, the limited number of samples included in the training sets for both methodologies (23 for Sanger sequencing and 39 for SNaPshot) should be determinant factors for the accuracy of the model.

In our study, comparing methylation levels obtained in both methodologies for the CpG site located at *ELOVL2* (Chr6:11044628), similar age-correlated values ($R = 0.823$ for Sanger sequencing; $R = 0.846$ for SNaPshot) were obtained, suggesting that both methodologies showed high accuracy in DNAm quantification.

When comparing the present DNAm data from buccal swabs to data obtained in blood samples from living Portuguese individuals captured by the same Sanger

sequencing and SNaPshot methodologies (as previously addressed in *Chapter 4. Results and discussion: A. DNA methylation age estimation in blood samples*), different methylation features can be found. Using Sanger sequencing, we observed that:

Page | 244

i) the best site from *ELOVL2* locus is different (CpG1 in buccal swabs vs. CpG6 in blood samples from living individuals);

ii) DNAm levels of CpG1, CpG4, CpG5 showed very strong age correlation values in blood samples from living individuals ($0.916 \leq R \leq 0.936$), while in buccal swabs, only strong age-correlated values were obtained for the same positions ($0.787 \leq R \leq 0.823$); this can revealed a specific age-association of each CpG in different tissue types;

iii) the MAD value obtained in buccal swabs with two CpGs from *ELOVL2* is higher (MAD = 8.32 years) (revealing lower accuracy) comparing with APM developed in blood samples from living individuals (MAD = 5.35 years) using four CpGs from *ELOVL2*, *EDARADD*, *PDE4C* and *FHL2*; however, it is known that the model accuracy could be improved by the inclusion of a larger number of CpGs or genes (Horvath, 2013; Lin *et al.*, 2016).

Regarding the SNaPshot methodology, the best age-correlated marker for buccal swabs was the *TRIM59* gene, while for blood samples was the *ELOVL2* ($R = 0.951$). This can explain the need of made a previous selection of tissue-specific CpGs/genes for each tissue type. Moreover, in our study, the MAD value obtained for buccal swabs with only one CpG is higher (6.73 years) (lower accuracy) comparing to the APM developed for blood samples from living Portuguese individuals (MAD = 4.25 years) that included three CpGs located at *ELOVL2*, *FHL2* and *C1orf132* genes. Once again, the addition of more predictor variables can be an important factor in model accuracy. In concordance with our data, Jung *et al.* (2019) reported different tissue-specific models for blood and buccal

swabs, revealing lower accuracy in buccal swabs (MAD = 3.174 in blood vs. MAD = 3.823 in buccal swabs).

In our study, DNAm levels captured by both methodologies in buccal swabs revealed lower accuracy in older ages. The increase in MAD values with aging, reflecting lower accuracy in age estimation, can be the reflection of individual alterations in the rate of methylation change that occurs with age, being slight in youths and accumulating with age. These differences in DNAm levels with the increase of age, reported by our data, are in concordance with previous studies across several tissues (Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015a, 2015b; Hamano *et al.*, 2016; Pfeifer *et al.*, 2020). Additionally, our reports showed non-significant sex DNAm changes in both methodologies for all the investigated CpG sites (P -value >0.05), in concordance with Bekaert *et al.* (2015b).

In conclusion, we developed two APMs for buccal swabs from living Portuguese individuals using two different methods for evaluation of DNAm levels. A two-CpGs APM combining *ELOVL2* CpG1 and *ELOVL2* CpG4 markers, analyzed through bisulfite sequencing, allowed to obtain a MAD of 8.32 years. Through SNaPshot, a MAD of 6.73 years was obtained using methylation information of one CpG located at *TRIM59* gene. We should note that the present study suffers from several limitations, being the small sample set the major drawback. Despite a larger sample set of buccal swabs was collected for the present study, the amount of extracted DNA was often insufficient for age-related DNAm analysis.

Chapter 4. Results and discussion

E. DNA methylation age estimation through multi-tissues

1. Introduction

A wide range of publications based on DNA methylation (DNAm) analysis for age estimation has emerged in the forensic field. Several age-related markers have been investigated in different tissues, including blood, saliva, buccal swabs, sperm, teeth, bones, allowing the development of tissue-specific age prediction models (APMs) with similar high accuracy in age prediction (Goel *et al.*, 2017). The discovery of universal biomarkers of age applied simultaneously to many tissue types can be a challenge, since it has been observed that only few markers can work well as multi-tissue age predictive markers (Jung *et al.*, 2019). However, the development of multi-tissue APMs can brought many advantages in forensics, since they can be applicable to several contexts with different types of samples.

Horvath (2013) assessed to methylation information of 353 CpGs developing a highly accurate multi-tissue age predictive model showing a strong correlation between predicted and chronological ages ($R = 0.97$), and revealing a median absolute difference between chronological and predicted ages of 2.9 years (training set) and 3.6 years (test set). The high accuracy can be explained by the larger number of CpGs included in the model. However, the need of a high number of age markers can also bring a challenge for forensic casework application. Moreover, in the Horvath model a larger error (around 10 years) was observed in several tissues suggesting that the best markers for one tissue may not be the best for another.

Eipel *et al.* (2016) reported that using a specific APM with methylation information of age-correlated markers selected in one tissue-specific type can lead to a decrease in model accuracy in age prediction if applied to a different tissue. This should be related with the tissue-specific differences in epigenetic patterns (Illingworth *et al.*, 2008; Li *et al.*, 2010; Bernstein *et al.*, 2010).

Jung and colleagues (2019) develop a multi-tissue APM with DNAm captured by SNaPshot assay from five CpGs located at *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes applied to blood, buccal swabs and saliva. The multi-tissue model showed high accuracy with a Mean Absolute Deviation (MAD) from chronological age of 3.553 years. This MAD value is similar to the reported by the same study when developing tissue-specific APMs (MAD = 3.17 years in blood; MAD = 3.82 years in buccal swabs; MAD = 3.29 years in saliva). In addition, Jung and colleagues (2019) have observed that *FHL2* gene is more tissue-specific revealing strong positive age correlation values in saliva and blood, and a weak age correlation in buccal swabs. They observed also that *ELOVL2* and *TRIM59* seem to work as better multi-tissue markers than *FHL2*, *C1orf132* or *KLF14*.

We previously assessed to the methylation information of age-correlated CpG sites in genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, *C1orf132*, *TRIM59* and *KLF14*, captured by Sanger sequencing and SNaPshot methodologies, using several tissue types including, blood, tooth, bone and buccal swabs, to develop tissue-specific APMs. In the present study, we reexamined DNAm levels of these highly age-correlated genes combining various tissues in order to test for multi-tissue APMs. For this main purpose, we: i) evaluated, by means of simple linear regression analysis, the age correlation value for single CpG sites in the combined set of samples; and ii) developed multi-locus multi-tissue APMs based simultaneously in the methylation information of several genes and/or CpGs.

2. Materials and Methods

2.1. Sample collection

A total of 245 samples collected from Portuguese individuals was used in this study for development of multi-tissue APMs, including: 71 blood samples from living individuals (45 females, 26 males; aged 1-95 years old); 73 blood samples from deceased individuals (15 females, 58 males; aged 24-91 years old); 39 buccal swabs from living individuals (16 males, 23 females; aged 3-86 years old); 31 tooth samples (10 males, 21 females; aged 26-94 years old; living individuals, n = 23; BDS, n = 8); and 31 bone samples collected from autopsies (26 males, 5 females; aged 26-81 years old).

We excluded methylation information captured from bones collected from BDS because these were subjected to an embalming treatment, which can have an effect in DNAm levels, as described in *Chapter 4. Results and Discussion: C. DNA methylation age estimation in fresh bone samples.*

2.2. DNAm analyses

Processing of samples (DNA extraction, quantification, bisulfite conversion, Sanger sequencing and SNaPshot assay) was made in accordance with previously described in *Chapter 3. Sample and design research.*

2.3. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). As previously reported in *Chapter 3. Sample and design research,* simple and multiple linear regression models

were used to analyze relationships between DNAm levels at CpG sites and chronological age for both Sanger sequencing and SNaPshot methodologies.

For Sanger sequencing data, three groups of combined samples were assigned as training sets:

Page | 252

- Group 1, addressing methylation information captured from CpGs located at *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes from all samples of blood, teeth and bones (**Supplementary Table S18**). We excluded buccal swabs because DNAm was only investigated in *ELOVL2*. Moreover, we did not include the methylation information captured from the *C1orf132* gene, because this locus was not investigated in blood samples from living individuals.
- Group 2, addressing methylation information of all genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* in blood samples from deceased individuals, teeth and bones (**Supplementary Table S19**). We excluded blood samples from living individuals (*C1orf132* was not studied) and buccal swabs (where only the *ELOVL2* was investigated).
- Group 3, specific for the nine CpG sites of the *ELOVL2* locus, comprising DNAm information in all tissue types (**Supplementary Table S20**).

For data obtained through the SNaPshot assay, two groups of combined tissues were used:

- Group 1, included methylation information of *ELOVL2*, *KLF14*, *TRIM59*, *C1orf132*, *FHL2* genes captured in blood samples from living and deceased individuals, tooth samples from living and deceased individuals and bone samples collected from autopsies;
- Group 2, included DNAm levels of the three CpGs located at *ELOVL2*, *KLF14* and *TRIM59* genes in all these tissues and buccal swabs, since in buccal swabs only these three CpG sites have been evaluated.

The stepwise regression analysis was used to select the best combination of age-associated sites in developing multi-locus APMs. For both methodologies, Spearman correlation coefficients and MAD values, between chronological and predicted ages, and the root mean square error (RMSE) were calculated.

3. Results

The association between chronological age and DNAm levels of CpGs located at *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132* was assessed in several types of samples by the bisulfite PCR sequencing methodology. Moreover, DNAm levels of CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were reexamined in the same tissues through the multiplex methylation SNaPshot assay, described by Jung *et al.* (2019). We built several groups of combined samples in each methodology for development of multi-tissue APMs. Some samples did not amplify for all the investigated genes in each group.

3.1. Multi-tissue APMs using Sanger sequencing

Using DNAm information from the *ELOVL2*, *EDARADD*, *PDE4C*, *C1orf132* and *FHL2* genes assessed by the Sanger sequencing methodology, we developed multi-tissue APMs using the three groups of combined samples previously detailed in the Materials and Methods section. In groups 1 and 2 we develop: i) simple linear regression APMs with the best age-associated CpG from each gene; and ii) multi-locus APMs with the selected CpGs from all genes using a stepwise regression approach. In group 3, we developed a simple linear regression model with the best CpG from *ELOVL2* and a final APM with the selected CpGs from *ELOVL2* gene by stepwise approach.

Correlation between DNAm levels and chronological age in group 1

Page | 254

Thirty-seven CpG sites (*ELOVL2*: 9 CpGs; *EDARADD*: 4 CpGs; *FHL2*: 12 CpGs; and *PDE4C*: 12 CpGs) were selected for methylation evaluation through the bisulfite PCR sequencing method in the training set of 204 samples, including blood (from living and deceased individuals), teeth (from living and deceased individuals) and bone collected during autopsies (85 females, 119 males; aged 1-94 years). The *ELOVL2* locus showed the highly significant age-correlation values for all the CpG sites ($R \geq 0.575$), reflecting the similar strength of change in DNAm with age across all CpGs. The first six CpGs located at *FHL2*, showed a $R \geq 0.512$ (**Supplementary Table S18**), however for the remaining sites the age correlation values revealed a negligible or weak correlation. For *EDARADD* CpG2 and *EDARADD* CpG3, a moderate age correlation was observed as well as for *PDE4C* CpG5 and *PDE4C* CpG1 to CpG3 (**Supplementary Table S18**).

Considering the strongest age-correlated site in each gene, the best marker was *ELOVL2* CpG5 with a correlation value between DNAm and chronological age of 0.706, (P -value = 1.12×10^{-31}), explaining 49.6% of the variation in age, followed by *EDARADD* CpG3 ($R = -0.682$, P -value = 5.27×10^{-29}), explaining 46.3% of the variation in age, *FHL2* CpG1 ($R = 0.662$; P -value = 5.92×10^{-27}), explaining 43.5% of the variation in age, and *PDE4C* CpG2 ($R = 0.605$; P -value = 4.50×10^{-21}), explaining 36.3% of the variation in age (**Table 4.54; Supplementary Table S18**). For these sites, a clear positive age correlation was observed for *ELOVL2* CpG5, *PDE4C* CpG2 and *FHL2* CpG1 markers, and a clear negative age correlation was observed for *EDARADD* CpG3 marker (**Supplementary Figure S32**).

The predicted age of individuals was calculated using the simple linear regression coefficients for the individual four strongest age-associated markers, allowing to obtain MAD values of 12.98 years for *ELOVL2* CpG5, 13.47 years for *EDARADD* CpG3, 13.72

years for *FHL2* CpG1 and 13.99 years for *PDE4C* CpG2 (**Table 4.54; Supplementary Figure S33**).

Table 4.54: Simple and multiple linear regression statistics of the age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes to test for association between the DNAm levels and chronological age in the group 1.

Locus	CpG site	Location	Multi-tissue: type of samples included	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>										
<i>ELOVL2</i>	CpG5	Chr6:11044642	Blood* + Bones + Teeth	201	0.706	0.499	0.496	16.24	1.12×10^{-31}	12.98
<i>EDARADD</i>	CpG3	Chr1:236394382	Blood* + Bones + Teeth	202	-0.682	0.465	0.463	16.64	5.27×10^{-29}	13.47
<i>PDE4C</i>	CpG2	Chr19:18233133	Blood* + Bones + Teeth	196	0.605	0.366	0.363	18.32	4.50×10^{-21}	13.99
<i>FHL2</i>	CpG1	Chr2:105399282	Blood* + Bones + Teeth	203	0.662	0.438	0.435	17.07	5.92×10^{-27}	13.72
<i>Multiple linear regression</i>										
APM (<i>ELOVL2</i> CpG5, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2, <i>PDE4C</i> CpG5, <i>PDE4C</i> CpG6, <i>PDE4C</i> CpG9, <i>FHL2</i> CpG1, <i>FHL2</i> CpG5 and <i>FHL2</i> CpG11)			Blood* + Bones + Teeth	194	0.932	0.868	0.862	8.51	1.97×10^{-76}	6.42

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly. *Blood samples from living and deceased individuals.

Testing simultaneously the 28 significant age-associated CpGs from *ELOVL2* (9 CpGs), *EDARADD* (3 CpGs), *FHL2* (9 CpGs) and *PDE4C* (7 CpGs) by stepwise regression analysis, a multi-locus multi-tissue APM that selected nine CpGs (*ELOVL2* CpG5, *EDARADD* CpG3, *PDE4C* CpG2, *PDE4C* CpG5, *PDE4C* CpG6, *PDE4C* CpG9, *FHL2* CpG1, *FHL2* CpG5 and *FHL2* CpG11) was constructed. This model revealed an age correlation value, $R = 0.932$ ($P\text{-value} = 1.97 \times 10^{-76}$), explaining 86.2% of the variation in age (corrected $R^2 = 0.862$) (**Table 4.54**). The final multi-locus multi-tissue APM developed in 194 samples allowed to predict age of individuals through the formula (**Table 4.55**): $32.920 + 51.682 \times \text{DNAm level } ELOVL2 \text{ CpG5} - 34.322 \times \text{DNAm level } EDARADD \text{ CpG3} + 34.419 \times \text{DNAm level } PDE4C \text{ CpG2} + 59.975 \times \text{DNAm level } PDE4C \text{ CpG5} - 41.886 \times \text{DNAm level } PDE4C \text{ CpG6} - 68.194 \times \text{DNAm level } PDE4C \text{ CpG9} - 85.035 \times \text{DNAm level } FHL2 \text{ CpG1} + 139.387 \times \text{DNAm level } FHL2 \text{ CpG5} - 58.241 \times \text{DNAm level } FHL2 \text{ CpG11}$. The obtained correlation between predicted and chronological ages was 0.902 (Spearman correlation coefficient) (**Figure 4.38**), with a MAD from chronological age of 6.42 years (RMSE = 8.26). Correct predictions were 75.7% assuming that chronological and predicted ages match ± 9 years, according to the standard error of estimate calculated for the final APM (SE = 8.51).

Table 4.55: Statistical parameters obtained in a multiple regression model with the nine CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, selected by stepwise regression approach, in blood, bone and tooth samples.

Marker	Coefficient	P-value
(Intercept)	32.920	0.004
<i>ELOVL2</i> CpG5	51.682	0.000
<i>FHL2</i> CpG1	-85.035	0.001
<i>FHL2</i> CpG5	139.387	0.000
<i>FHL2</i> CpG11	-58.241	0.000
<i>EDARADD</i> CpG3	-34.322	0.002
<i>PDE4C</i> CpG2	34.419	0.003
<i>PDE4C</i> CpG5	59.975	0.000
<i>PDE4C</i> CpG6	-41.886	0.000
<i>PDE4C</i> CpG9	-68.194	0.000

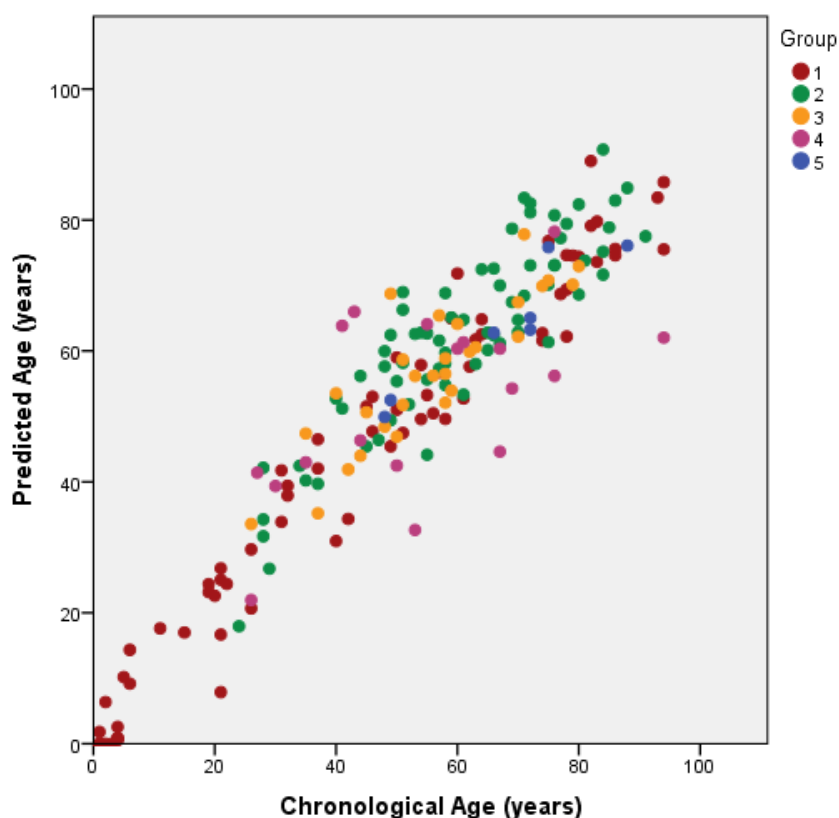


Figure 4.38: Predicted age *versus* chronological age using the multi-locus multi-tissue APM developed for *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4) and tooth samples from deceased individuals (5).

Correlation between DNAm levels and chronological age in group 2

Forty-three CpG sites (*ELOVL2*: 9 CpGs; *EDARADD*: 4 CpGs; *FHL2*: 12 CpGs, *PDE4C*: 12 CpGs; and *C1orf132*: 6 CpGs) were selected for methylation evaluation through the bisulfite PCR sequencing method in the training set of 133 samples, including blood (from deceased individuals), teeth (from living and deceased individuals) and bone collected during autopsies (40 females, 93 males; aged 24-94 years).

All the CpGs located at *ELOVL2* locus showed significant age correlation values ($0.425 \leq R \leq 0.563$) (**Supplementary Table S19**). In general, for all CpGs from the remaining genes *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* we observed a lower age association comparing with *ELOVL2* CpGs (**Supplementary Table S19**).

Testing the correlation between methylation levels and chronological age, the highest age-correlated site was *ELOVL2* CpG5 ($R = 0.563$; $P\text{-value} = 3.09 \times 10^{-12}$), explaining 31.2% of the variation in age. For the remaining genes, the CpG from each locus with the strongest age-correlated value was *PDE4C* CpG2 ($R = 0.470$; $P\text{-value} = 2.87 \times 10^{-8}$), explaining 21.4% of the variation in age, *C1orf132* CpG1 ($R = -0.455$, $P\text{-value} = 7.04 \times 10^{-8}$), explaining 20% of the variation in age, *EDARADD* CpG3 ($R = -0.416$, $P\text{-value} = 7.90 \times 10^{-7}$), explaining 16.6% of the variation in age, and *FHL2* CpG1 ($R = 0.365$; $P\text{-value} = 0.000017$), explaining 12.7% of the variation in age (**Table 4.56; Supplementary Table S19**). A clear positive correlation was observed for *ELOVL2* CpG5, *PDE4C* CpG2 and *FHL2* CpG1, and a clear negative age correlation for *EDARADD* CpG3 and *C1orf132* CpG1 (**Supplementary Figure S34**). Predicting age based on DNAm levels of these sites through simple linear regression allowed to obtain MAD values of 10.97 years for *ELOVL2* CpG5, 11.36 years for *C1orf132* CpG1, 11.87 years for *PDE4C* CpG2, 12.40 years for *EDARADD* CpG3 and 12.64 years for *FHL2* CpG1 (**Table 4.56; Supplementary Figure S35**).

Table 4.56: Simple and multiple linear regression statistics of the age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between the DNAm levels and chronological age in the group 2.

Locus	CpG site	Location	Multi-tissue: type of samples included	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>										
<i>ELOVL2</i>	CpG5	Chr6:11044642	Blood** + Bones + Teeth	130	0.563	0.317	0.312	13.84	3.09×10^{-12}	10.97
<i>C1orf132</i>	CpG1	Chr1:207823681	Blood** + Bones + Teeth	126	-0.455	0.207	0.200	14.67	7.04×10^{-8}	11.36
<i>PDE4C</i>	CpG2	Chr19:18233133	Blood** + Bones + Teeth	125	0.470	0.221	0.214	14.70	2.87×10^{-8}	11.87
<i>EDARADD</i>	CpG3	Chr1:236394382	Blood** + Bones + Teeth	131	-0.416	0.173	0.166	15.03	7.90×10^{-7}	12.40
<i>FHL2</i>	CpG1	Chr2:105399282	Blood** + Bones + Teeth	132	0.365	0.133	0.127	15.35	0.000017	12.64
<i>Multiple linear regression</i>										
APM (<i>C1orf132</i> CpG4, <i>FHL2</i> CpG5, <i>FHL2</i> CpG6, <i>FHL2</i> CpG8 and <i>ELOVL2</i> CpG5)			Blood** + Bones + Teeth	123	0.797	0.635	0.620	10.09	2.71×10^{-24}	7.27

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly. **Blood samples from deceased individuals only (*C1orf132* was not studied in blood samples from living individuals using Sanger sequencing methodology).

Evaluating simultaneously the 30 significant age-associated CpGs (among the 43 CpGs from *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132* genes) by stepwise regression approach, we developed a multi-locus multi-tissue APM with methylation information of *FHL2* CpG5, *FHL2* CpG6, *FHL2* CpG8, *ELOVL2* CpG5 and *C1orf132* CpG4. This multi-locus APM showed an age correlation value $R = 0.797$ (P -value = 2.71×10^{-24}) and explained 62% of the variation in age (corrected $R^2 = 0.620$) (**Table 4.56**). The predicted age of each individual was calculated using the multivariate regression coefficients through the formula (**Table 4.57**): $17.244 - 53.968 \times \text{DNAm level } C1orf132 \text{ CpG4} + 156.170 \times \text{DNAm level } FHL2 \text{ CpG5} - 82.244 \times \text{DNAm level } FHL2 \text{ CpG6} - 61.144 \times \text{DNAm level } FHL2 \text{ CpG8} + 76.131 \times \text{DNAm level } ELOVL2 \text{ CpG5}$. The model showed a MAD from chronological age of 7.27 years (RMSE = 9.68) (**Table 4.56**), and a correlation between predicted and chronological ages (**Figure 4.39**) of 0.799 (Spearman correlation coefficient). Correct predictions were 77.2% assuming that chronological and predicted ages match ± 10 years, according to the standard error of estimate calculated for the final APM (SE = 10.09).

Table 4.57: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2* and *C1orf132*, selected by stepwise regression approach, in blood, bone and tooth samples.

Marker	Coefficient	P-value
(Intercept)	17.244	0.127
<i>FHL2</i> CpG5	156.170	0.000
<i>FHL2</i> CpG6	-82.244	0.046
<i>FHL2</i> CpG8	-61.144	0.004
<i>ELOVL2</i> CpG5	76.131	0.000
<i>C1orf132</i> CpG4	-53.968	0.000

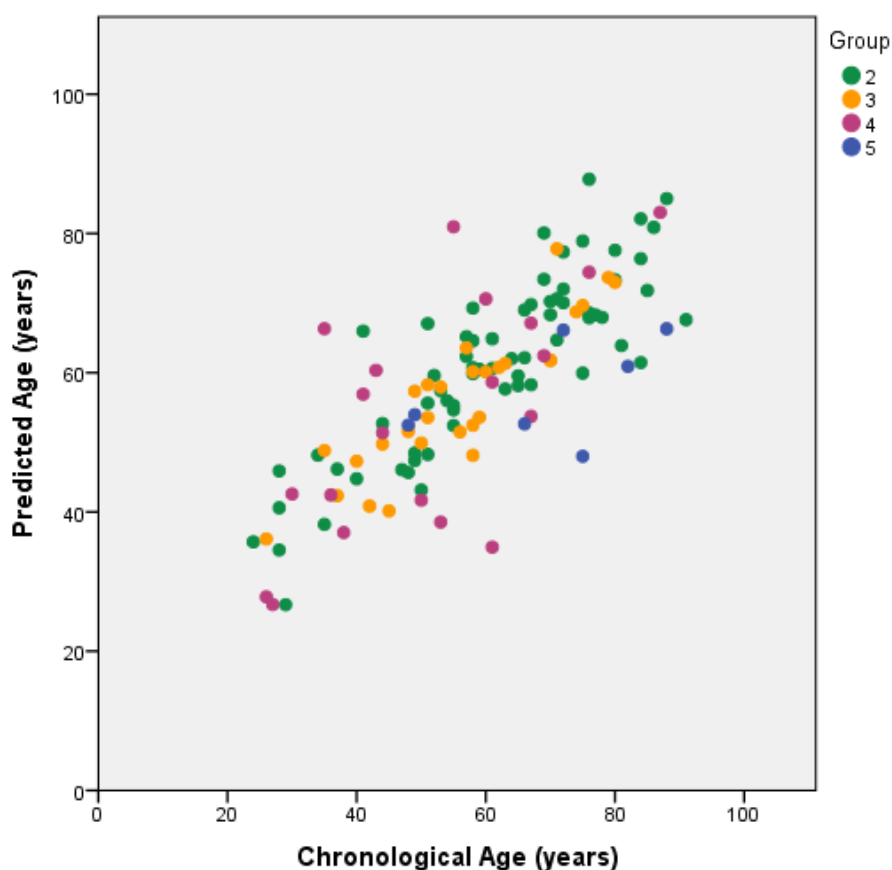


Figure 4.39: Predicted age *versus* chronological age using the multi-tissue model developed for *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes including blood samples from deceased individuals (2), bone samples (3) and tooth samples from living (4) and deceased individuals (5).

Correlation between DNAm levels and chronological age in group 3

Methylation information of nine CpGs located at *ELOVL2* gene were reexamined in a training set of 227 samples including all the tissue types (blood, buccal swabs, teeth and bones). A moderate or strong age correlation values ($0.582 \leq R \leq 0.736$) was observed for all the CpGs (**Supplementary Table S20**). The strongest age-correlated value was observed for *ELOVL2* CpG5 ($R = 0.736$; $P\text{-value} = 2.07 \times 10^{-39}$), explaining 53.9% of the variation in age (**Table 4.58; Supplementary Table S20**). Predicted age using this site allowed to obtain a MAD of 13.07 years (**Table 4.58; Supplementary Figure S36**).

DNAm information of simultaneously all CpGs tested through the stepwise regression allowed the selection of CpG3, CpG5, CpG8 and CpG9 for development of a

final linear regression model. Applying this APM, the age prediction for each individual was obtained through the formula detailed in **Table 4.58**. The obtained correlation between predicted and chronological age was 0.752 (Spearman correlation coefficient), the MAD from chronological age was 12.09 years (**Table 4.58; Supplementary Figure S37**), and the rate of correct predictions was 67.4% considering the cutoff of ± 15 years, according to the standard error of estimate calculated for the final APM (SE = 15.38).

Table 4.58: Multi-tissue APM based in *ELOVL2* methylation information obtained using Sanger sequencing methodology.

Linear regression analysis (CpGs selected)	Multi-tissue: type of samples included	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>								
APM (DNAm level <i>ELOVL2</i> CpG5, Chr6:11044642)	Blood* + Bones + Teeth + Buccal swabs	224	0.736	0.541	0.539	16.47	2.07×10^{-39}	13.07
<i>Multiple linear regression</i>								
APM (Predicted Age = $(-70.085) + 106.833 \times \text{DNAm level } ELOVL2 \text{ CpG3} + 75.533 \times \text{DNAm level } ELOVL2 \text{ CpG5} + 80.890 \times \text{DNAm level } ELOVL2 \text{ CpG8} - 98.512 \times \text{DNAm level } ELOVL2 \text{ CpG9}$)	Blood* + Bones + Teeth + Buccal swabs	224	0.778	0.605	0.598	15.38	4.26×10^{-43}	12.09

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic position was based on the GRCh38/hg38 assembly. *Blood samples from living and deceased individuals.

3.2. Multi-tissue APMs using SNaPshot methodology

DNAm information of five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes captured by SNaPshot methodology, was used to develop multi-tissue APMs in two groups of combined samples, as previously described in *Materials and Methods*.

Correlation between DNAm levels and chronological age in group 1

DNAm levels at five CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were simultaneously measured through a SNaPshot assay in the combined set of 176 samples, including blood (living and deceased individuals), bones and teeth from Portuguese individuals (105 males, 71 females; 1-94 aged years old). DNAm levels of *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes revealed a positive correlation with chronological age, and DNAm levels of *C1orf132* showed a negative age correlation (**Supplementary Figure S38**).

Testing the individual DNAm association with chronological age for the five CpG sites, the strongest correlation was observed for *ELOVL2* ($R = 0.779$, $P\text{-value} = 3.17 \times 10^{-36}$), explaining 60.4% of the variation in age, followed by *FHL2* ($R = 0.697$, $P\text{-value} = 9.21 \times 10^{-27}$), explaining 48.3% of the variation in age, *KLF14* ($R = 0.683$, $P\text{-value} = 8.37 \times 10^{-25}$), explaining 46.3 % of the variation in age, *C1orf132* ($R = -0.682$, $P\text{-value} = 4.04 \times 10^{-25}$), explaining 46.2% of the variation in age, and *TRIM59* ($R = 0.595$, $P\text{-value} = 3.66 \times 10^{-18}$), explaining 35.1% of the variation in age (**Table 4.59**). Simple APMs for each CpG site allowed to obtain MAD values from chronological age of 10.97 years for *ELOVL2*, 12.46 years for *C1orf132*, 12.64 years for *KLF14*, 12.71 years for *FHL2* and 13.85 years for *TRIM59* (**Table 4.59**; **Supplementary Figure S39**).

Applying the stepwise regression approach, only the CpGs located at *ELOVL2*, *KLF14* and *C1orf132* genes were selected for the development of a final multi-locus APM. The multi-tissue APM with *ELOVL2*, *KLF14* and *C1orf132* markers showed a very strong age correlation, $R = 0.922$ ($P\text{-value} = 3.14 \times 10^{-67}$), explaining 84.7% of the variation in age (corrected $R^2 = 0.847$) (**Table 4.59**). Predicting age through the multivariate regression coefficients as follows (**Table 4.60**): $29.220 + 96.850 \times \text{DNAm level } ELOVL2 + 208.747 \times \text{DNAm level } KLF14 - 33.437 \times \text{DNAm level } C1orf132$, allowed to obtain a MAD from chronological age of 6.49 years (RMSE = 8.42) (**Table 4.59**). Correct predictions were 73.8% considering the cutoff of ± 9 years, according to the standard error of estimate calculated for the final APM (SE = 8.53). The obtained Spearman correlation value between predicted and chronological ages was 0.893 (**Figure 4.40**).

Table 4.59: Simple and multiple linear regression statistics at the five CpGs of the *ELOVL2*, *FHL2*, *KLF14*, *TRIM59* and *C1orf132* genes using SNaPshot assay in the group 1.

Locus	Location	Multi-tissue: type of samples included	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	Chr6:11044628	Blood* + Bones + Teeth	172	0.779	0.606	0.604	13.908	3.17×10^{-36}	10.97
<i>FHL2</i>	Chr2:105399282	Blood* + Bones + Teeth	175	0.697	0.486	0.483	15.902	9.21×10^{-27}	12.71
<i>KLF14</i>	Chr7:130734355	Blood* + Bones + Teeth	171	0.683	0.466	0.463	15.991	8.37×10^{-25}	12.64
<i>C1orf132</i>	Chr1:207823681	Blood* + Bones + Teeth	174	-0.682	0.465	0.462	16.286	4.04×10^{-25}	12.46
<i>TRIM59</i>	Chr3:160450189	Blood* + Bones + Teeth	175	0.595	0.354	0.351	17.855	3.66×10^{-18}	13.85
<i>Multiple linear regression</i>									
APM (<i>ELOVL2</i> , <i>KLF14</i> and <i>C1orf132</i>)		Blood* + Bones + Teeth	168	0.922	0.850	0.847	8.53	3.14×10^{-67}	6.49

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.*Blood samples from living and deceased individuals.

Table 4.60: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *C1orf132* and *KLF14*, selected by stepwise regression approach, in blood, bone and tooth samples.

Marker	Coefficient	P-value
(Intercept)	29.220	0.000
<i>C1orf132</i>	-33.437	0.000
<i>ELOVL2</i>	96.850	0.000
<i>KLF14</i>	208.747	0.000

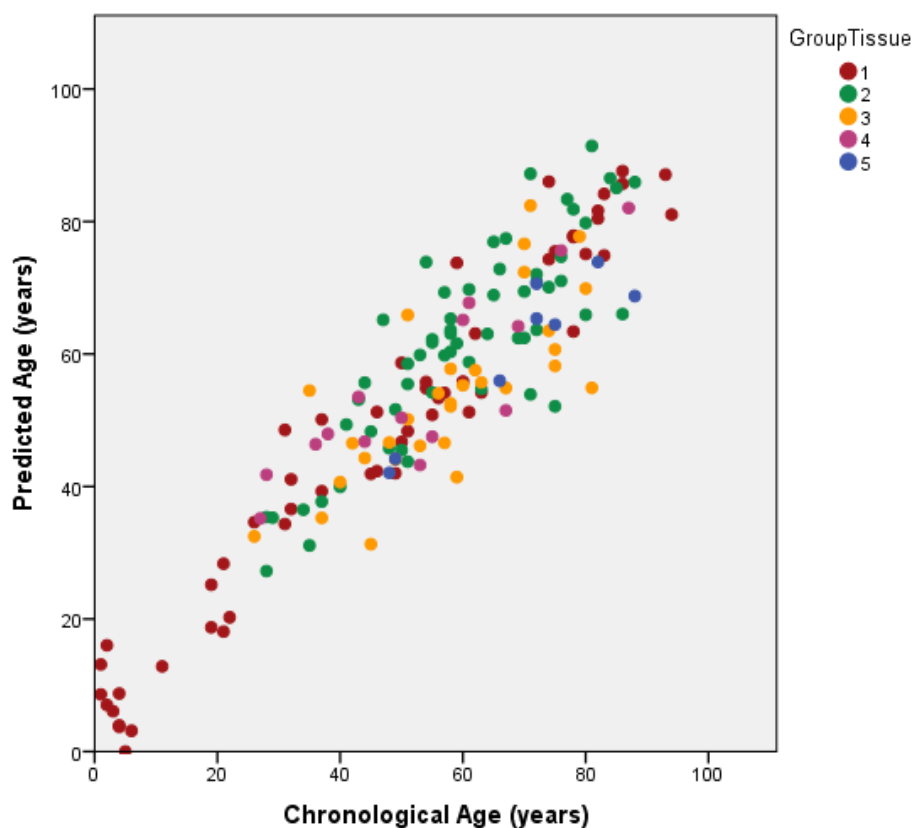


Figure 4.40: Predicted age *versus* chronological age using the multi-tissue APM developed for *ELOVL2*, *C1orf132* and *KLF14* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4) and tooth samples from deceased individuals (5).

Correlation between DNAm levels and chronological age in the group 2

The DNAm levels of three CpG sites located at *ELOVL2*, *KLF14* and *TRIM59* genes, were simultaneously measured through the SNaPshot assay in a combined set of

215 samples that included blood, bones, teeth and also buccal swabs (121 males, 94 females; aged 1-94 years old).

The simple linear regression analysis allowed to obtain positive correlations between DNAm and chronological age for all these genes (**Supplementary Figure S40**). The strongest age correlation value was observed for *ELOVL2* ($R = 0.793$, $P\text{-value} = 3.95 \times 10^{-46}$), explaining 62.6% of the variation in age. The remaining two genes revealed lower age correlation values, *KLF14* ($R = 0.477$, $P\text{-value} = 1.44 \times 10^{-30}$) and *TRIM59* ($R = 0.323$, $P\text{-value} = 2.29 \times 10^{-19}$), explaining 47.5% and 32.0% of the variation in age, respectively (**Table 4.61**). Predicted age for each CpG site revealed MAD values from chronological age of 11.40 years for *ELOVL2*, 13.93 years for *KLF14* and 16.43 years for *TRIM59* (**Table 4.61**; **Supplementary Figure S41**).

Based on the stepwise regression approach, a multi-tissue APM with *ELOVL2* and *KLF14* genes was developed, revealing a strong age correlation value, R , of 0.870 ($P\text{-value} = 1.95 \times 10^{-61}$) and explaining 75.4% of the variation in age (corrected $R^2 = 0.754$) (**Table 4.61**). Predicted age using the multiple regression equation (**Table 4.62**): $(-2.832) + 120.282 \times \text{DNAm level } ELOVL2 + 268.955 \times \text{DNAm level } KLF14$, allowed to obtain a MAD of 9.02 years (RMSE = 11.85) and a rate of correct predictions of 71.6% (standard error of estimate, SE = 11.95). Predicted and chronological ages revealed a correlation value (Spearman correlation coefficient, r) of 0.840 (**Figure 4.41**).

Table 4.61: Simple and multiple linear regression statistics at the three CpGs of the *ELOVL2*, *KLF14* and *TRIM59* genes using SNaPshot assay in the group 2.

Locus	Location	Multi-tissue: type of samples included	N	R	R ²	Corrected R ²	SE	P-value	MAD
<i>Simple linear regression</i>									
<i>ELOVL2</i>	Chr6:11044628	Blood* + Bones + Teeth + Buccal swabs	208	0.793	0.628	0.626	14.84	3.95×10^{-46}	11.40
<i>KLF14</i>	Chr7:130734355	Blood* + Bones + Teeth + Buccal swabs	206	0.691	0.477	0.475	17.41	1.44×10^{-30}	13.93
<i>TRIM59</i>	Chr3:160450189	Blood* + Bones + Teeth + Buccal swabs	210	0.568	0.323	0.320	19.98	2.29×10^{-19}	16.43
<i>Multiple linear regression</i>									
APM (<i>ELOVL2</i> and <i>KLF14</i>)		Blood* + Bones + Teeth + Buccal swabs	201	0.870	0.756	0.754	11.95	1.95×10^{-61}	9.02

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.* Blood from living and deceased individuals.

Table 4.62: Statistical parameters obtained in a multiple regression model with the two CpGs in genes *ELOVL2* and *KLF14*, selected by stepwise regression approach, in blood, bone, tooth and buccal swab samples.

Marker	Coefficient	P-value
(Intercept)	-2.832	0.225
<i>ELOVL2</i>	120.282	0.000
<i>KLF14</i>	268.955	0.000

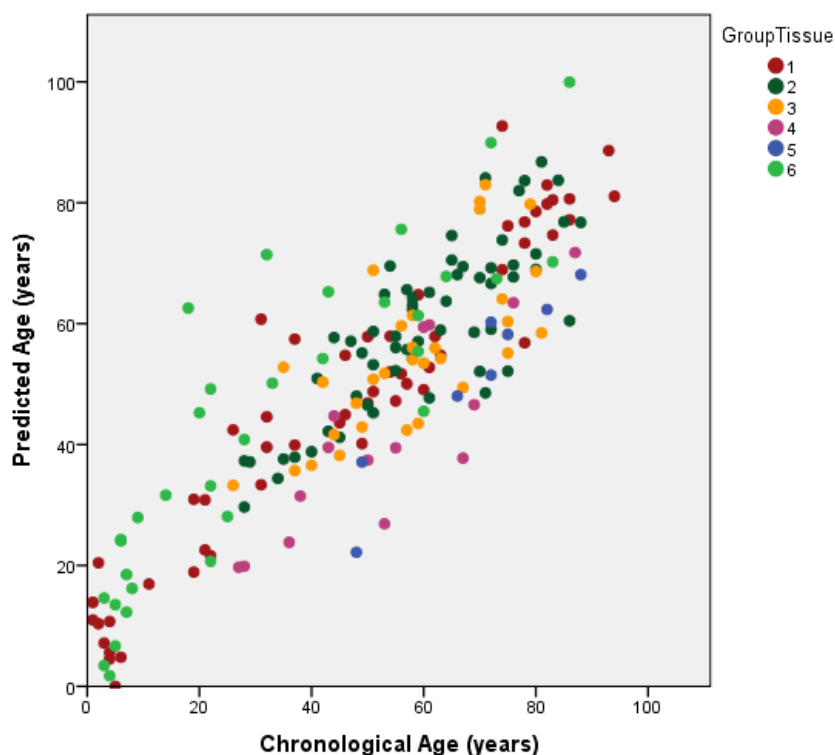


Figure 4.41: Predicted age *versus* chronological age using the multi-tissue APM developed for *ELOVL2*, *KLF14* and *TRIM59* genes including blood samples from living individuals (1), blood samples from deceased individuals (2), bone samples (3), tooth samples from living individuals (4), tooth samples from deceased individuals (5) and buccal swabs (6).

4. Discussion

Specific APMs with high accuracy have been developed using many tissue types in the past decade (e.g. Bocklandt *et al.*, 2011; Garagnani *et al.*, 2012; Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015b; Eipel *et al.*, 2016; Naue *et al.*, 2018; Gopalan *et al.*,

2019; Márquez-Ruiz *et al.*, 2020). However, the discovery of DNAm age-related markers with similarly high accuracy in different types of tissues (universal markers) is a challenging task in the forensic field (Koch and Wagner, 2011). In fact, only few studies have developed multi-tissue APMs (Horvath, 2013, Alsaleh *et al.*, 2017; Jung *et al.*, 2019). Instead, each age-correlated marker seems to reveal a specific ability to predict chronological age, as each can be affected by different intrinsic or environmental factors. A careful selection of age-associated CpGs and the validation of these proposed markers in each tissue type should be the first step for the development of multi-tissue APMs.

We reexamined DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* genes, previously captured in different tissue types (blood samples from living and deceased individuals; tooth samples; fresh bone samples; and buccal swabs) by Sanger sequencing and SNaPshot methodologies to build tissue-specific APMs, and, in the present study, we developed multi-tissue APMs combining sets of these tissues. In each group of combined tissues (three groups made with DNAm data captured by Sanger sequencing and two groups with DNAm levels captured by SNaPshot) we developed simple linear regression APMs for the best-selected CpG sites from each gene, and multi-locus APMs using the best combination of CpGs selected by the stepwise regression approach.

In all the three groups of combined samples evaluated through Sanger sequencing methodology, we observed that *ELOVL2* CpG5 was the strongest site (group 1, $R = 0.706$; group 2, $R = 0.563$; group 3, $R = 0.736$). This position revealed the most accurate MAD value in their respective groups 1, 2 and 3 (MAD = 12.98 years; MAD = 10.97 years; MAD = 13.07 years, respectively).

The stepwise approach combining the significant age-correlated CpGs from group 1 (28 CpGs from four genes) and from group 2 (30 CpGs from five genes), allowed to

develop multi-locus multi-tissue APMs with nine CpGs in group 1 and with five CpGs in group 2. The multi-locus multi-tissue APM developed in group 1 (*ELOVL2* CpG5, *EDARADD* CpG3, *PDE4C* CpG2, *PDE4C* CpG5, *PDE4C* CpG6, *PDE4C* CpG9, *FHL2* CpG1, *FHL2* CpG5 and *FHL2* CpG11) revealed a very strong age correlation value ($R = 0.932$), highly significant ($P\text{-value} = 81.97 \times 10^{-76}$), explaining 86.2% of the variation in age. This model developed with 194 Portuguese individuals (aged 1-95 years old), included blood samples (living and deceased), tooth samples (living and deceased) and bone samples collected during autopsies, and allow to predict age with moderate accuracy showing a MAD from chronological age of 6.42 years. The second multi-locus multi-tissue APM developed in group 2 by Sanger sequencing (*C1orf132* CpG4, *FHL2* CpG5, *FHL2* CpG6, *FHL2* CpG8 and *ELOVL2* CpG5) revealed a lower age correlation value ($R = 0.797$; $P\text{-value} = 2.71 \times 10^{-24}$), explaining 62% of the variation in age, allowing to obtain also a moderate accurate MAD value of 7.27 years. This model was developed without blood samples from living individuals and could be less informative. However, as we observed previously in this study (*Chapter 4. Results and discussion: A. DNA methylation age estimation in blood samples, 3.4. Applicability of the developed APMs for blood samples from living and deceased individuals*), APMs developed for blood samples from deceased individuals can be applied to blood samples from living individuals with similar accuracy. Consequently, we can hypothesize that the developed APM without methylation information captured in blood samples from living individuals can be applied to blood samples from living individuals with similar MAD value. Meanwhile, this should be accepted with caution and need to be further tested. Other relevant aspect, is that the percentage of the variation in age explained by the first multi-locus APM (developed in the group 1) is very high (86.2%), and can reflect the presence of more CpGs in the model (nine CpGs) and/or the larger size of the sample set.

We should note that the models developed in groups 1 and 2 included CpGs from *ELOVL2* and *FHL2* genes, revealing that these markers can be promising age-associated genes for the development of multi-tissue APMs. It has been shown that *ELOVL2* is a stable epigenetic marker that can reveal higher age correlation value across several tissues (Hannum *et al.*, 2013; Gopalan *et al.*, 2019). In recent years, this locus has been used as a powerful age-correlated marker in many APMs developed for blood, tooth, bones and buccal swabs, revealing similar high accuracy in all APMs (Garagnani *et al.*, 2012; Weidner *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Xu *et al.*, 2015; Hamano *et al.*, 2016; Giuliani *et al.*, 2016; Freire-Aradas *et al.*, 2016; Cho *et al.*, 2017; Thong *et al.*, 2017; Naue *et al.*, 2018; Gopalan *et al.*, 2019; Jung *et al.*, 2019; Márquez-Ruiz *et al.*, 2020). Moreover, it has been demonstrated that *ELOVL2* can be a high-performance multi-tissue marker for age predictions (Slieker *et al.*, 2018; Naue *et al.*, 2018; Jung *et al.*, 2019). The *FHL2* gene has already been used in several studies in blood, tooth and buccal swabs (Garagnani *et al.*, 2012; Zbieć-Piekarska *et al.*, 2015a; Hamano *et al.*, 2016; Giuliani *et al.*, 2016; Freire-Aradas *et al.*, 2016; Cho *et al.*, 2017; Thong *et al.*, 2017; Jung *et al.*, 2019) and more recently tested in bone samples (Lee *et al.*, 2020). Meanwhile, as defended by Jung and collaborators (2019) this marker and the *C1orf132* seem to have tissue-specific characteristics. Particularly for *FHL2* gene, a strong age correlation value was observed for blood and saliva samples, while weak age correlation was obtained for buccal swabs (Jung *et al.*, 2019).

The multi-tissue model developed in group 3 with *ELOVL2* CpG3, CpG5, CpG8 and CpG9 sites, can be applied to blood (from living and deceased individuals), teeth, bones and buccal swabs, revealing a MAD value of 12.09 years. The advantage of this model is the inclusion of buccal swabs but, at the same time, only included CpGs located at *ELOVL2* gene. As this model showed a higher MAD comparing with the previous

multi-locus APMs developed in groups 1 and 2, and considering also the tissue-specific APM developed exclusively for buccal swabs (MAD = 8.32 years) (see *Chapter 4. Results and discussion: D. DNA methylation age estimation in buccal swabs*), it does not seem very useful for age predictions in forensic contexts.

Regarding methylation information captured by SNaPshot methodology, the best CpG was the position located at *ELOVL2* gene both in group 1 (blood, bones and teeth), $R = 0.779$, and in group 2 (blood, bones, teeth and buccal swabs), $R = 0.628$. In group 1, the multi-locus multi-tissue APM obtained by stepwise regression addressing methylation information of five CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes, allowed the selection of three sites located at *ELOVL2*, *KLF14* and *C1orf132* genes. This model revealed a very strong age correlation value ($R = 0.922$), with a MAD from chronological age of 6.49 years. In the group 2 (which included buccal swabs), DNAm levels of the three sites located at *ELOVL2*, *KLF14* and *TRIM59* genes allowed the development of a multi-locus multi-tissue model with *ELOVL2* and *KLF14* genes. This dual-locus model revealed a strong age correlation ($R = 0.870$) and a MAD from the chronological age of 9.02 years. In concordance with data reported using Sanger sequencing, in both multi-locus APMs developed using the SNaPshot methodology, the CpG from *ELOVL2* was selected, revealing, once again, the powerful of this marker for development of multi-tissue APMs in forensic contexts.

In conclusion, in this study we reexamined DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* genes captured by Sanger sequencing and SNaPshot methodologies, and several multi-tissue APMs were developed using blood, buccal swabs, teeth and bones samples from Portuguese individuals. By Sanger sequencing, we obtained an accurate MAD of 6.42 years in the multi-locus APM developed for blood (from living and deceased individuals), bones and tooth samples

using nine CpGs from four different genes (*ELOVL2*, *FHL2*, *PDE4C* and *EDARADD*). A similar MAD of 7.27 years was obtained in the multi-locus APM developed with blood (from deceased individuals), bone and tooth samples using five CpGs from three different genes (*ELOVL2*, *FHL2* and *C1orf132*). Using the SNaPshot assay, a multi-locus APM was developed combining *C1orf132*, *ELOVL2* and *KLF14* genes in a multi-tissue sample with blood (from living and deceased individuals), bone and teeth, revealing a MAD from chronological age of 6.49 years. The multi-locus APM more suitable for a multi-tissue sample combining blood (from living and deceased individuals), bone, tooth and buccal swabs included genes *ELOVL2* and *KLF14*, revealing a MAD of 9.02 years. Both methodologies revealed similar accuracy for use in multi-tissue APMs being both simple, rapid and cost-effective and easily available in forensic laboratories. The multi-tissue models herein developed can be a promising tool for age estimation in forensic contexts.

Chapter 4. Results and discussion

F. Dry bone samples from *Coleção de Esqueletos Identificados do Século XXI - CEI/XXI*

1. Introduction

Forensic anthropologists have to deal with human identification, not only in routine practice, but also in exceptional scenarios, associated with mass disaster or crimes against humanity including war crimes and human rights violation (de Boer *et al.*, 2019; de Boer *et al.*, 2020). Age estimation represents one of the big four parameters within the biological profile for identification purposes. Meanwhile, age estimation remains one of the most difficult task in forensic contexts. For estimating age in forensic anthropology, the selection of appropriate methods depends on the type of bones present, the state of preservation of these bones, and the age range in which the person is included. Nowadays, several anthropological methods to estimate age are available for fetal, children, adults, and older adults based on different skeletal or dental indicators (Adserias-Garriga and Wilson-Taylor, 2019; Cunningham, 2019).

Age estimation based on DNA methylation (DNAm) has emerged as a new tool in the forensic field for the development of age prediction models (APMs). Most forensic studies addressing DNAm age prediction were performed across different tissues, including blood, buccal swabs or saliva, however few studies have focused on bone samples (Horvath *et al.*, 2015, 2018; Naue *et al.*, 2018; Gopalan *et al.*, 2019; Lee *et al.*, 2020). Moreover, these studies used fresh bones collected from patients (Horvath *et al.*, 2015, 2018; Gopalan *et al.*, 2019), fresh bone samples collected during autopsy (Naue *et al.*, 2018) or collected from Bodies Donated to Science (Gopalan *et al.*, 2019). Moreover, in the study of Gopalan *et al.* (2019), they also used forensic bone samples with 2-3 years of natural decomposition in an outdoor environment, which have been excluded due to non-reliable results (Gopalan *et al.*, 2019).

This study aimed to estimate age of skeletonized individuals of *Coleção de Esqueletos Identificados do Século XXI* (CEI/XXI), which have different *postmortem* interval (PMI) comparing with the previously reported studies, through DNAm analysis.

2. Materials and Methods

2.1. Sample collection

The 21st Century Identified Skeletal Collection (ISC/XXI) or *Coleção de Esqueletos Identificados do Século XXI* (CEI/XXI), housed in the Laboratory of Forensic Anthropology, Department of Life Sciences at the University of Coimbra, arises from a collaboration in 2007 between the Department of Anthropology of the University of Coimbra and the City Council of *Santarém* (responsible for the cemetery from which the skeletons originate) to carry the unclaimed skeletons to the University for investigation purposes. At this moment (2021), CEI/XXI has around 302 skeletons almost exclusively Portuguese individuals, which died between 1982 and 2012. The collection has been used in many investigations related with different forensic anthropology issues as ancestry, PMI, age-at-death, analysis of burned human remains (Ferreira *et al.*, 2020). For age estimation investigation, the main drawback of the CEI/XXI is the advanced age of the individuals, which limits forensic investigations (Ferreira *et al.*, 2014). Particularly, only 12.25% of the individuals aged less than 61 years (Ferreira *et al.*, 2020).

A set of nine femur of skeletonized individuals from the CEI/XXI (7 females, 2 males; aged 38-92 years old) was selected (**Table 4.63**). As DNA extraction from bones requires fragmentation of the sample, being a destructive process, a careful choice of skeletons from CEI/XXI was made to avoid the degradation of the collection. Moreover,

a small number of samples was first selected to control the cost and time consumed of DNA analysis. We considered two key factors for the selection:

- i) state of preservation of the skeleton, mainly of the bone in question (femur);
- ii) age of the individual;

All the skeletons selected have until 18 years of PMI, which represents the time between the death and the analysis of the human remains (**Table 4.63**). A brief description (in Portuguese) about the features, state of preservation, *postmortem* conditions and measurements of selected bone was made in Annex VII.

Table 4.63: *Antemortem* data of selected skeletons from CEI/XXI.

ID	Age	Sex	Death date	Collection and analysis date
CEI/XXI 87	83 years	Female	05/04/2000	2018
CEI/XXI 128	38 years	Female	03/12/2008	2018
CEI/XXI 62	60 years	Male	07/06/2000	2018
CEI/XXI 103	92 years	Female	05/01/2007	2018
CEI/XXI 122	55 years	Male	06/01/2008	2018
CEI/XXI 123	83 years	Female	08/02/2007	2018
CEI/XXI 267	59 years	Female	14/07/2004	2018
CEI/XXI 164	48 years	Female	08/07/2008	2019
CEI/XXI 262	58 years	Female	16/01/2011	2019

2.2. Bone processing

The processing of dry bone samples (femur) was made at INMLCF, following the standard guidelines, as previously described in *Chapter 3. Sample and design research*. Briefly, we made a pre-treatment of bone sample, which consists in cleaning the bone (with bleach and then distilled water) and grinding it, in a room designated exclusively for processing human remains.

Two additional pre-incubations with ethylene diamine tetra acetic acid (EDTA) were performed at LGH from CIAS, before DNA extraction (Protocol 1 – EDTA and

Protocol 2 – EDTA). The treatment with EDTA was made to improve the amount of DNA extracted from bones (Zupanič Pajnič *et al.*, 2016).

The Protocol 1 – EDTA, adapted from the *User-Developed Protocol: Purification of total DNA from compact animal bone using the DNeasy® Blood & Tissue Kit* (Qiagen, 2006), consisted in: incubation of the bone powder (100-150 mg) with 5 ml EDTA (Promega) 0.5 M for 24 hours at room temperature on a rotator mixer (**Figure 4.42**). After centrifugation for 5 minutes at 2000 g, the EDTA solution was discarded. These steps were repeated for 4-5 days. Then, about 150 mg of bone powder was used for DNA extraction.

The Protocol 2 – EDTA consisted in: a pre-incubation of 100 mg of bone powder with 2 ml EDTA 0.5 M, pH 8 for 1h; after centrifugation for 5 minutes at 2000 g, the EDTA solution was discarded. Then only 50 mg of bone powder was placed in a lysis tube for DNA extraction.

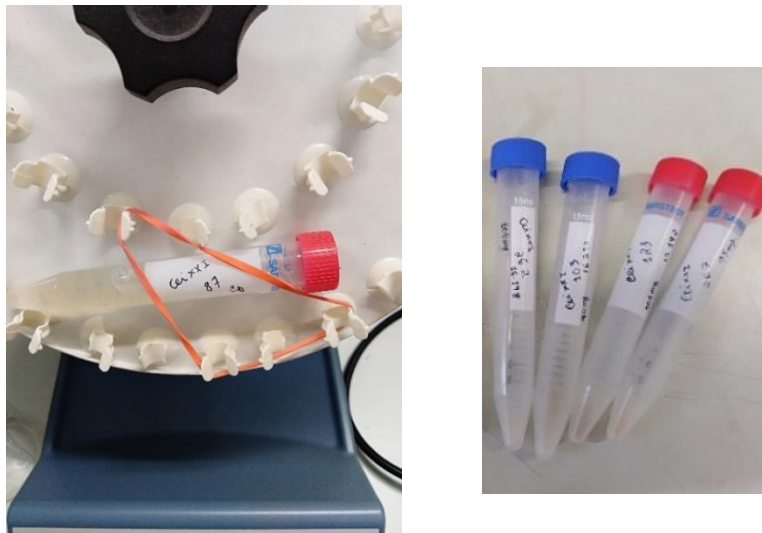


Figure 4.42: A) Incubation of bone powder of skeleton CEI/XXI 87 with Protocol 1 – EDTA. B) Bone powder of several skeletons form CEI/XXI (around 100-150 mg) before the incubation with EDTA.

2.3. DNA extraction, quantification and bisulfite conversion

Subsequent DNA extraction and purification were performed at the INMLCF using *PrepFiler Express BTA™ Forensic DNA extraction kit* (Applied Biosystems, Foster City, CA) as described in *Chapter 3. Sample and design research*. For Protocol 1- EDTA bone samples, a modified protocol was used, which consists in triplicate the volume of reagents (this means each sample requires the following: 660 µl PrepFiler BTA™ lysis Buffer, 9 µl freshly prepared 1 M DTT, 21 µl Proteinase K). For Protocol 2- EDTA bone samples, the standard protocol, previously described in *Chapter 3. Sample and design research*, was used.

DNA Quantification was performed with the real-time PCR-based *Quantifiler™ Human DNA Quantification Kit* (Applied Biosystems, Foster City, CA) at the INMLCF, as described in *Chapter 3. Sample and design research*.

After that, genomic DNA was subjected to bisulfite conversion using the *EZ DNA Methylation-Gold™ Kit* (Zymo Research, Irvine, USA) according to the instructions of manufacturer, as previously described in *Chapter 3. Sample and design research*. Briefly, for bones from CEI/XXI, 30 µl of genomic DNA (in a total amount of 0.063 ng to 0.216 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 µl.

2.4. Polymerase chain reaction (PCR) and Sanger sequencing

After bisulfite conversion, the modified DNA samples were submitted to PCR for the selected regions of genes *ELOVL2*, *FHL2* and *C1orf132* using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described

in *Chapter 3. Sample and design research*. For bones from CEI/XXI, 1.5 μ l of the bisulfite converted DNA was used in the total PCR volume of 25 μ l.

The design research used for each sample is presented in **Table 4.64**.

Table 4.64: Design research made using bones from CEI/XXI.

ID	Bone preparing	Normal/Standard Protocol			Pre-treatment with EDTA		
		DNA extraction	DNA quantification	Bisulfite conversion, PCR, sequencing	DNA extraction	DNA quantification	Bisulfite conversion, PCR, sequencing
CEI/XXI 87	✓	✓	✓	✓	Protocol 1 – EDTA	✓	✓
CEI/XXI 128	✓	✓	✓	-	Protocol 1 – EDTA	✓	-
CEI/XXI 62	✓	-	-	-	Protocol 2 – EDTA	✓	✓
CEI/XXI 103	✓	-	-	-	Protocol 2 – EDTA	✓	✓
CEI/XXI 122	✓	-	-	-	Protocol 1 – EDTA	-	-
CEI/XXI 123	✓	-	-	-	-	-	-
CEI/XXI 267	✓	-	-	-	Protocol 2 – EDTA	✓	-
CEI/XXI 164	✓	-	-	-	-	-	-
CEI/XXI 262	✓	-	-	-	Protocol 2 – EDTA	✓	-

3. Results

In the present study, we used bone samples from the CEI/XXI to extract DNA.

Page | 286

Despite the extracted amount of DNA could not be enough for a successful DNAm analysis, we performed bisulfite conversion and PCR-sequencing for *ELOVL2*, *FHL2* and *C1orf132* genes in few samples from the collection.

Standard protocol

First, we selected an older individual (CEI/XXI 87, aged 83 years old) and a younger individual (CEI/XXI 128, aged 38 years old) for DNA extraction through the standard protocol, routinely used at the INMLCF (protocol without EDTA, as previously described in *Chapter 3. Sample and design research*). As observed in **Table 4.65**, for CEI/XXI 87, both VICTM dye, which represents the internal PCR control (IPC) detector, and FAMTM dye, which represents the Quantifiler Human detector, had amplification (positive PCR result), despite the very low amount of DNA (0.0035 ng/ μ l). For the CEI/XXI 128, the VICTM dye (IPC detector) had amplification, but the amplification of the Quantifiler Human detector did not occur; this means that there is a successful amplification for IPC target, but the human-specific target was not amplified. There is a true negative result and consequently, there is no DNA in the sample (**Table 4.65**).

Table 4.65: Quantification data obtained in dry bones from CEI/XXI with the standard protocol.

ID	Protocol	Quantity Human	Quantity IPC	Quantity ng/ μ l	Interpretation
CEI/XXI 87	Without EDTA	36.93	27.98	0.0035	Positive result
CEI/XXI 128	Without EDTA	Undetermined	27.88	-	Negative Result

For the skeleton CEI/XXI 87, we proceeded with the DNAm analysis, including bisulfite conversion and PCR amplification for CpGs located at *ELOVL2*, but we obtained a complete failure of PCR amplification.

Protocol 1 – EDTA

Then, we used an alternative protocol with a pre-incubation with EDTA in these two selected skeletons (Protocol 1 – EDTA), but the amount of the extracted DNA remained very low (**Table 4.66**).

Table 4.66: Quantification data obtained in dry bones from CEI/XXI with the alternative Protocol 1 – EDTA.

ID	Protocol	Quantity Human	Quantity IPC	Quantity ng/μl	Interpretation
CEI/XXI 87	Protocol 1- EDTA	35.90	27.48	0.004679	Positive result
CEI/XXI 128	Protocol 1- EDTA	36.97	27.68	0.002087	Positive result

As the skeleton CEI/XXI 87 revealed a higher amount of DNA, we performed a new bisulfite conversion, amplification and sequencing for the *ELOVL2* and *C1orf132* genes. We obtained a reduced PCR performance in both PCR reactions, however the sequencing chromatogram allowed to assess to the DNAm levels (**Table 4.67**).

Table 4.67: Methylation quantification of CpGs located at *ELOVL2* and *C1orf132* genes using the femur sample of the skeleton CEI/XXI 87.

CEI/XXI 87	<i>ELOVL2</i> C/(C+T)	CEI/XXI 87	<i>C1orf132</i> C/(C+T)
CpG1	0.215726	CpG1	0
CpG2	0.21131	CpG2	0.08237
CpG3	0.234907	CpG3	0
CpG4	0.221178	CpG4	0
CpG5	0.228864	CpG5	0
CpG6	0.092448	CpG6	0.081006
CpG7	0.204248		
CpG8	0.280985		
CpG9	0.29793		

Protocol 2 – EDTA

Using another protocol with EDTA (Protocol 2 – EDTA) and selecting four different skeletons (CEI/XXI 103, aged 92 years old; CEI/XXI 62, aged 60 years old; CEI/XXI 262, aged 58 years old; CEI/XXI 267, aged 59 years old), similar lower amounts of DNA were obtained (**Table 4.68**).

Table 4.68: Quantification data obtained in dry bones from CEI/XXI with the alternative Protocol 2 – EDTA.

ID	Protocol	Quantity Human	Quantity IPC	Quantity ng/μl	Interpretation
CEI/XXI 103	Protocol 2 – EDTA	36.10	27.65	0.004012	Positive result
CEI/XXI 62	Protocol 2 – EDTA	35.32	27.53	0.007233	Positive result
CEI/XXI 262	Protocol 2 – EDTA	36.80	27.03	0.002385	Positive result
CEI/XXI 267	Protocol 2 – EDTA	36.46	27.23	0.003073	Positive result

Despite the low DNA concentration, we have performed the bisulfite conversion and PCR amplification in skeletons CEI/XXI 62 and CEI/XXI 103. Bisulfite converted

DNA samples showed a complete failure of PCR amplification or a reduced PCR performance. The Sanger sequencing for the *FHL2* gene in these PCR products showed no reliable results in the sequencing chromatograms. Hence, no further analysis was made with this set of bones.

4. Discussion

We addressed skeletonized individuals from the CEI/XXI for DNAm analysis using bisulfite conversion and PCR sequencing methodology.

Three different protocols for DNA extraction were used: i) Standard protocol, previously described in *Chapter 3. Sample and design research*, without EDTA and using 50 mg of bone powder in DNA extraction); ii) Protocol 1 – EDTA (using a pre-incubation with EDTA, and around 150 mg of bone powder + EDTA which required 3x of the volume of reagents, being more expensive); and iii) Protocol 2 – EDTA (with EDTA, and 50 mg of bone powder + EDTA). The amount of extracted DNA was low in all the processed skeletons, with the protocols with EDTA-treatment giving similar performances to the obtained with the standard protocol without EDTA. Nevertheless, we observed a little improvement of the amount of extracted DNA with the EDTA pre-incubation.

We obtained DNA concentrations in the extractions between 0.0021 ng/μl and 0.0071 ng/μl that is a very low amount of DNA to be subjected to bisulfite conversion, in which occurs DNA degradation or fragmentation through the conversion. The bone samples from CEI/XXI analyzed in this study revealed a complete failure in PCR or reduced performance in amplification after bisulfite conversion, excepting for the skeleton CEI/XXI 87 in which methylation data of CpGs located at *ELOVL2* and

C1orf132 genes were observed. Nevertheless, we must note that this bone sample belongs to an older individual (83 years old), with putative high patterns of DNAm for these two genes. However, comparing with DNAm data obtained in other bone samples collected from individuals during autopsies and from BDS with the same chronological age or aged between 79-87 years old (previously described in *Chapter 4. Results and discussion: C. DNA methylation age estimation in fresh bone samples*), we observed that no consistent and reliable DNAm data were obtained for the skeletonized individual CEI/XXI 87. Indeed, in bones collected during autopsies and/or from BDS, the methylation level is higher than 50% in all the CpGs located in both *C1orf132* and *ELOVL2* genes, while for CEI/XXI 87 we observed values lower than 29% for all the CpGs (as shown in **Table 4.67**). This can be related with the state of preservation of the skeleton as result of environmental *postmortem* conditions, which leads to degradation of DNA in bone samples. Our samples from CEI/XXI can be considered “ancient samples” not due the PMI (of about 18 years, **Table 4.63**) but due the very low amount and degradation of the DNA extracted from bones. We hypothesize that the bisulfite PCR sequencing used in our analysis do not have sufficient sensitivity to address a low quantity of DNA or degraded DNA obtained from these types of samples. Other reports employing massively parallel sequencing (MPS) have shown successful results with limited amount and/or highly degraded DNA. Pedersen *et al.* (2014) assessed to DNAm levels of permafrost hair samples collected from a Paleo-Eskimo with 4000 years old, and predicted age-at-death. This reveals the reliability of assess to DNAm levels and predict age based on methylation data from ancient samples using more accurate techniques such as MPS.

Moreover, it has been observed that ancient DNA can suffer *postmortem* miscoding lesions (DDMLs), which cause incorrect incorporation of nucleotides, as

deamination (Hofreiter *et al.* 2001; Brotherton *et al.*, 2007). *Postmortem* deamination is a spontaneously chemical process that occurs due the hydrolytic deamination of cytosine residues into uracils (Hofreiter *et al.* 2001). This can affect the rate of DNAm analysis through bisulfite conversion because if deamination occurs all the residues C (either unmethylated or methylated) can be transformed in T. As in our methylation data from bones from CEI/XXI we observed very low rate of DNAm, we can hypothesize that our samples also suffered from a deamination process. Despite of this phenomenon, it has been showed the stability of C^m patterns in ancient DNA, when preserved ancient DNA samples were used (Briggs *et al.*, 2010; Llamas *et al.*, 2012).

In conclusion, the amount of extracted DNA from a number of CEI/XXI skeletons was not sufficient for a subsequent successful DNAm analysis for age estimation purposes that include procedures as bisulfite conversion and successful PCR for Sanger sequencing. It is important to note that selected bones from CEI/XXI showed several states of conservation and have suffer from different *postmortem* influences, which can influence the amount of DNA present in bones.

Meanwhile, we should not discard the possibility of using DNA recovered from skeletonized individuals from CEI/XXI in forensic science studies. As an example, the bone powder of these nine skeletonized individuals was used in another research study in the field of forensics, where similar amounts of extracted DNA allowed the successful analysis of ancestry-informative markers (AIMs) (Botelho, 2020).

The use of skeletonized individuals from CEI/XXI, with several states of preservation and *postmortem* influences, was a challenge in terms of DNAm analysis, at least in this work. Despite the different procedures used, the obtained quantity of DNA extracted from skeletonized human remains did not allow successful results in the analysis of DNAm for age estimation purposes.

Chapter 5. General discussion

Chapter 5. General discussion

DNA methylation (DNAm) is currently one of the hottest topics in forensics, being accepted as a promising tool for age estimation. In recent years, several DNAm-derived epigenetic clocks have been proposed using several types of biological tissues, different techniques and many CpGs (Bocklandt *et al.*, 2011; Garagnani *et al.*, 2012; Hannum *et al.*, 2013; Horvath, 2013; Florath *et al.*, 2014; Weidner *et al.*, 2014; Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015a, 2015b; Xu *et al.*, 2015; Lee *et al.*, 2015; Giuliani *et al.*, 2016; Eipel *et al.*, 2016; Park *et al.*, 2016; Freire-Aradas *et al.*, 2016, 2018; Cho *et al.*, 2017; Thong *et al.*, 2017; Spólnicka *et al.*, 2017; Hamano *et al.*, 2016, 2017; Naue *et al.*, 2017, 2018; Vidaki *et al.*, 2017; Hong *et al.*, 2017; Aliferi *et al.*, 2018; Naue *et al.*, 2018; Hong *et al.*, 2019; Jung *et al.*, 2019; Gopalan *et al.*, 2019; Daunay *et al.*, 2019; Lee *et al.*, 2020; Márquez-Ruiz *et al.*, 2020; Pfeifer *et al.*, 2020).

In the present study, DNAm levels of a number of CpG sites located in seven genes, *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, *C1orf132*, *KLF14* and *TRIM59*, were assessed by Sanger sequencing and SNaPshot methodologies for development of tissue-specific APMs using blood (*Chapter 4. Results and discussion: A. DNA methylation age estimation in blood samples*), teeth (*Chapter 4. Results and discussion: B. DNA methylation age estimation in tooth samples*), bones (*Chapter 4. Results and discussion: C. DNA methylation age estimation in fresh bone samples*) and buccal swabs (*Chapter 4. Results and discussion: D. DNA methylation age estimation in buccal swabs*). Moreover, using methylation information captured in these tissues, we developed multi-tissue APMs in combined sets of these different tissue types (*Chapter 4. Results and discussion: E.*

DNA methylation age estimation through multi-tissues). We have tested DNA extraction from skeletonized individuals for DNAm analysis in order to build APMs in these valuable sources of samples in forensics, however the amount of extracted DNA was not enough to obtain successful results (*Chapter 4. Results and discussion: F. Dry bone samples from Coleção de Esqueletos Identificados do Século XXI-CEI/XXI*). Additionally, we evaluated the effect of the *postmortem* interval (PMI) in DNAm levels in bones from BDS (*Chapter 4. Results and discussion: C. DNA methylation age estimation in fresh bone samples*).

Bisulfite Sanger sequencing was the first technique to be described for analyzing DNAm patterns. In this study, we analyzed a number of CpG sites by direct PCR sequencing of five previously known age-related genes, *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132*. Moreover, using the SNaPshot methodology, we assessed DNAm levels of five highly CpGs located at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C* and *TRIM59*, replicating the experiments proposed by Jung *et al.* (2019).

Considering methylation information of the highly age-associated markers in each tissue (**Table 5.1 and Table 5.2**), we observed that for both Sanger sequencing and SNaPshot methodologies, the best age-predictive sites in blood samples from living individuals were located at *ELOVL2* and *FHL2*, showing similar age-correlation results, whereas in blood samples from deceased individuals the best CpG site was located at *ELOVL2*. For bone samples from autopsies, DNAm levels captured by Sanger sequencing revealed the best age-associated site located at *ELOVL2*, while through the SNaPshot assay the best CpG site was located at *FHL2*. Bone samples from BDS revealed contrasting data across the several analyzed genes: for Sanger sequencing, the best marker

was located at *PDE4C*, and for SNaPshot all the genes revealed no significant age-correlated values. In tooth samples, the best marker seems to be the *FHL2* using Sanger sequencing, however SNaPshot revealed *KLF14* gene as the best marker. For buccal swabs, *TRIM59* was the best marker using the SNaPshot methodology, being *ELOVL2* the second best age-related predictor; however, only three genes were investigated through SNaPshot (*ELOVL2*, *KLF14* and *TRIM59*) in buccal swabs and only the *ELOVL2* gene was investigated by Sanger sequencing.

Considering the two methodologies used in this study, most CpGs show strong ($0.70 \leq R \leq 0.90$) or very strong ($R \geq 0.90$) correlation values in blood samples from living individuals (**Tables 5.1 and 5.2**). As the selection of candidate age-predictive markers for our study was made considering the higher accuracy obtained in previous reports, mainly focusing in blood samples from living individuals (Garagnani *et al.*, 2012; Florath *et al.*, 2014; Weidner *et al.*, 2014; Xu *et al.*, 2015; Zbieć-Piekarska *et al.*, 2015b; Cho *et al.*, 2017; Alghanim *et al.*, 2017; Alsaleh *et al.*, 2017; Jung *et al.*, 2019), it was expected that these markers may work better in this same tissue type.

Table 5.1: Age-correlated values for the best CpG sites in each tissue type using Sanger sequencing.

Gene	CpG site	Location	Blood living	Blood deceased	Blood living and deceased	Bone autopsies	Bone BDS	Teeth living and deceased	Buccal swabs
			R	R	R	R	R	R	R
<i>ELOVL2</i>	CpG1	Chr6:11044628	0.920	0.781	0.872	0.736	-0.223	0.237	0.823
	CpG3	Chr6:11044634	0.884	0.740	0.859	0.764	-0.006	0.379	0.631
	CpG4	Chr6:11044640	0.916	0.785	0.880	0.780	-0.077	0.221	0.787
	CpG6	Chr6:11044644	0.936	0.764	0.892	0.852	-0.059	0.336	0.683
<i>FHL2</i>	CpG1	Chr2:105399282	0.937	0.431	0.828	0.692	0.580	0.451	
	CpG2	Chr2:105399288	0.940	0.465	0.821	0.654	0.641	0.409	
	CpG3	Chr2:105399291	0.940	0.459	0.813	0.480	0.655	0.517	
	CpG4	Chr2:105399297	0.814	0.141	0.623	0.371	0.370	0.658	
<i>EDARADD</i>	CpG3	Chr1:236394382	-0.888	-0.621	-0.786	-0.561	-0.430	-0.124	
<i>PDE4C</i>	CpG1	Chr19:18233139	0.849	0.401	0.785	0.592	0.259	0.474	
	CpG2	Chr19:18233133	0.852	0.592	0.830	0.690	0.372	0.466	
	CpG3	Chr19:18233131	0.827	0.585	0.813	0.620	0.771	0.391	
<i>Clorf132</i>	CpG1	Chr1:207823681		-0.634		-0.834	-0.305	-0.123	

Abbreviations: R, correlation coefficient. The best CpG site from each gene in each tissue type is in bold. Genomic positions were based on the GRCh38/hg38 assembly.

Table 5.2: Age-correlated values obtained for the CpGs located at genes *ELOVL2*, *FHL2*, *KLF14*, *TRIM59* and *C1orf132* in each tissue type using the SNaPshot assay.

Gene	CpG site	Location	Blood living	Blood deceased	Blood living and deceased	Bone autopsies	Bone BDS	Teeth living and deceased	Buccal swabs
			R	R	R	R	R	R	R
<i>ELOVL2</i>	CpG1	Chr6:11044628	0.951	0.791	0.919	0.619	0.412	0.685	0.846
<i>FHL2</i>	CpG1	Chr2:105399282	0.946	0.654	0.874	0.708	0.245	0.331	
<i>KLF14</i>	CpG1	Chr7:130734355	0.791	0.568	0.731	0.540	0.256	0.728	0.821
<i>C1orf132</i>	CpG1	Chr1:207823681	-0.924	-0.591	-0.834	-0.507	-0.040	-0.080	
<i>TRIM59</i>	CpG1	Chr3:160450189	0.910	0.769	0.830	0.633	0.147	0.665	0.946

Abbreviations: R, correlation coefficient. The best CpG site in each tissue is in bold. Genomic positions were based on the GRCh38/hg38 assembly.

Simple linear regression was used to derive APMs for each tissue type using the best age-associated site in each gene, obtained through Sanger sequencing, and using each one of the five CpGs investigated by the SNaPshot assay. The observed values of MAD between predicted and chronological ages ranged between 5.53 years and 15.22 years, showing higher to lower accuracy. Using Sanger sequencing, the high age-prediction accuracy was obtained for *PDE4C* CpG3 (MAD = 5.53 years) in bones from BDS, followed by *ELOVL2* CpG6 (MAD = 5.73 years) in bone samples from autopsies, *FHL2* CpG3 (MAD = 7.81 years) in blood samples from living individuals, *ELOVL2* CpG6 (MAD = 8.76 years) in blood from living and deceased individuals, *ELOVL2* CpG4 (MAD = 8.89 years) in blood samples from deceased individuals, and *FHL2* CpG4 (MAD = 11.35 years) in tooth samples. For methylation data captured by SNaPshot, the best age-predictor reported in each tissue was also different: *ELOVL2* for blood samples from living and deceased individuals and for the overall training set of blood samples (MAD = 6.73, 7.64 and 7.46 years, respectively), *FHL2* for bone samples from autopsies (MAD = 7.95 years) and *KLF14* for teeth (MAD = 9.68 years). These results suggest that each age-associated marker could reveal specificity for each tissue type. Hence, the most suitable approach for accurate age predictions in specific tissues is the selection of specific markers showing the best age-correlation performance. It has been shown in previous reports that several genes and/or CpGs perform better in specific tissues (Weidner *et al.*, 2014; Eipel *et al.*, 2016). However, it seems that some markers can reveal good results across different tissues and can be used for the development of multi-tissue APMs with a similar accuracy (Horvath, 2013; Jung *et al.*, 2019).

It is well described that simple linear regression models for age prediction showed higher MAD values comparing with multi-locus APMs that include a variable number of

genes/CpGs. This means that the inclusion of more predictive variables from the same gene or from different genes can improve the model accuracy. Considering the multi-locus APMs developed in this study (**Table 5.3**), we observed that:

i) for blood samples from living individuals, the best APM included methylation information of *ELOVL2*, *C1orf132* and *FHL2* genes captured by the SNaPshot assay. This model revealed high accuracy with a MAD = 4.25 years. The model included less CpGs (three CpGs) comparing with that developed using Sanger sequencing (with four CpGs allowing a higher MAD value = 5.35 years).

ii) for blood samples from deceased individuals, the APM developed through SNaPshot methodology, with *ELOVL2*, *C1orf132*, *FHL2* and *TRIM59* genes, revealed a higher accuracy (MAD = 5.36 years) in comparison with the APM developed using Sanger sequencing methodology (MAD = 6.08 years). Once again, the model developed through SNaPshot methodology included less predictor variables (four CpGs) than the APM developed with methylation information captured by Sanger sequencing (five CpGs).

iii) the best multi-locus APM developed for tooth samples included methylation information of genes *ELOVL2* and *KLF14* captured by the SNaPshot assay. This model revealed a moderate accuracy (MAD = 7.07 years). Using Sanger sequencing in teeth, only CpGs located at *FHL2* gene revealed moderate or strong age correlation values, and it was not possible to develop a multi-locus model.

iv) for bones collected during autopsies, the best APM was developed with DNAm levels of *ELOVL2* CpG5, *ELOVL2* CpG7, *ELOVL2* CpG6, *C1orf132* CpG1, *EDARADD* CpG3 and *EDARADD* CpG4 markers captured by Sanger sequencing, revealing a high accuracy with a MAD of 2.56 years. Of note, the model included more predictor variables

(six CpGs) than the model developed with the methylation information captured by SNaPshot (two CpGs, MAD = 7.18 years), allowing to obtain a higher accuracy.

v) for buccal swabs, the dual-locus model developed by Sanger sequencing with CpGs from *ELOVL2* showed a MAD = 8.32 years. For SNaPshot methodology, it was not possible to develop a multi-locus model.

vi) for combined training set of blood samples from living and deceased individuals, the multi-locus APM (with *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes) developed with methylation information captured by SNaPshot revealed the higher accuracy (MAD = 4.97 years) in comparison with the model developed using Sanger sequencing with *EDARADD* CpG3, *FHL2* CpG1, *ELOVL2* CpG6 and *PDE4C* CpG2 markers (MAD = 6.21 years).

From these results, we can say that not only the gene/CpG selected or the number of CpGs included in the APM contribute to the model accuracy, but also the methodology used in evaluation of DNAm levels (**Table 5.3**). Usually, when we used more CpGs in the development of APMs, the model accuracy increases, as it has been observed in point iv. However, as it has been observed in point i, ii and v, the models developed through SNaPshot assay, combining less CpGs, can reveal high accuracy comparing to models developed by Sanger sequencing with more CpGs.

Table 5.3: Summary of the best tissue-specific APMs based on DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* captured by Sanger sequencing and SNaPshot methodologies.

Tissue	Method	APM	R	R ²	P-value	MAD	MAD Cross validation
Blood (Living)	Sanger sequencing	<i>ELOVL2</i> CpG6, <i>FHL2</i> CpG3, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2	0.972	0.946 (corrected R ² = 0.941)	1.11 × 10 ⁻²⁹	5.35	6.20*
	SNaPshot	<i>ELOVL2</i> , <i>FHL2</i> , <i>C1orf132</i>	0.982	0.965 (corrected R ² = 0.963)	7.32 × 10 ⁻³⁸	4.25	4.75*
Blood (Deceased)	Sanger sequencing	<i>ELOVL2</i> CpG4, <i>FHL2</i> CpG2, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2, <i>C1orf132</i> CpG1	0.888	0.788 (corrected R ² = 0.763)	8.17 × 10 ⁻¹³	6.08	7.22*
	SNaPshot	<i>ELOVL2</i> , <i>FHL2</i> , <i>C1orf132</i> , <i>TRIM59</i>	0.899	0.808 (corrected R ² = 0.793)	1.07 × 10 ⁻¹⁸	5.36	6.13*
Blood (Living and Deceased)	Sanger sequencing	<i>ELOVL2</i> CpG6, <i>FHL2</i> CpG1, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2	0.947	0.897 (corrected R ² = 0.894)	3.470 × 10 ⁻⁶⁶	6.21	6.42*
	SNaPshot	<i>ELOVL2</i> , <i>FHL2</i> , <i>C1orf132</i> , <i>TRIM59</i>	0.963	0.928 (corrected R ² = 0.925)	1.03 × 10 ⁻⁶¹	4.97	5.25*
Bones (Autopsies)	Sanger sequencing	<i>ELOVL2</i> CpG5, <i>ELOVL2</i> CpG6, <i>ELOVL2</i> CpG7, <i>EDARADD</i> CpG3, <i>EDARADD</i> CpG4, <i>C1orf132</i> CpG1	0.970	0.941 (corrected R ² = 0.925)	2.097 × 10 ⁻¹²	2.56	3.77**
	SNaPshot	<i>FHL2</i> , <i>KLF14</i>	0.777	0.604 (corrected R ² = 0.576)	0.000002	7.18	7.84**
Bones (BDS)	Sanger sequencing	<i>FHL2</i> CpG2, <i>PDE4C</i> CpG3	0.851	0.725 (corrected R ² = 0.694)	0.000009	4.67	6.39**
Teeth (Living and Deceased)	Sanger sequencing	<i>FHL2</i> CpG4***	0.658	0.433 (corrected R ² = 0.413)	0.000078	11.35	12.22**
	SNaPshot	<i>ELOVL2</i> , <i>KLF14</i>	0.886	0.785 (corrected R ² = 0.764)	2.09 × 10 ⁻⁷	7.07	7.33**

Buccal swabs (Living)	Sanger sequencing	<i>ELOVL2</i> CpG1, <i>ELOVL2</i> CpG4***	0.894	0.800 (corrected $R^2 = 0.780$)	1.035×10^{-7}	8.32	11.80**
	SNaPshot	<i>TRIM59</i> ***	0.946	0.894 (corrected $R^2 = 0.891$)	1.183×10^{-17}	6.73	7.07**

Abbreviations: R, correlation coefficient. MAD, mean absolute deviation (years) between chronological and predicted ages.

* 4 fold;

** 3 fold;

*** APM built by simple linear regression analysis with the best CpG or by multiple linear regression with CpGs from the same locus.

Considering the multi-locus multi-tissue APMs developed in this study, combining sets of samples from the different tissues under study (**Table 5.4**), the best model (MAD = 6.49 years) was developed in group 1 (including blood from living and deceased individuals, bone from autopsies and tooth samples), with three CpGs located at *ELOVL2*, *KLF14* and *C1orf132* genes captured by the SNaPshot assay. Despite, this value is slightly higher than the obtained in the model developed by Sanger sequencing (nine CpGs, MAD = 6.42 years). However, the SNaPshot model has only three CpGs, being more useful for forensic casework.

Combining samples of blood from living and deceased individuals, bones from autopsies, teeth, and buccal swabs (group 3 in Sanger sequencing and group 2 in SNaPshot), the best multi-locus multi-tissue APM was the model developed with methylation information captured by SNaPshot (APM with *ELOVL2* and *KLF14*; MAD = 9.02 years) (**Table 5.4**). We should note that in the training set of buccal swabs only the *ELOVL2* gene was evaluated using Sanger sequencing (in group 3) and *ELOVL2*, *KLF14* and *TRIM59* genes using SNaPshot (in group 2). Thus, we assessed to DNAm levels of less age-correlated markers for development of all the APMs.

Table 5.4: Summary of the multi-locus multi-tissue APMs based on DNAm levels of *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD*, *C1orf132*, *TRIM59* and *KLF14* captured by Sanger sequencing and SNaPshot methodologies.

Tissues	Group	Method	APM	R	R ²	P-value	MAD
Blood* + Bones autopsies + Teeth	Group 1	Sanger sequencing	<i>ELOVL2</i> CpG5, <i>EDARADD</i> CpG3, <i>PDE4C</i> CpG2, <i>PDE4C</i> CpG5, <i>PDE4C</i> CpG6, <i>PDE4C</i> CpG9, <i>FHL2</i> CpG1, <i>FHL2</i> CpG5, <i>FHL2</i> CpG11	0.932	0.868 (corrected R ² = 0.862)	1.97 × 10 ⁻⁷⁶	6.42
	Group 1	SNaPshot	<i>ELOVL2</i> , <i>KLF14</i> , <i>C1orf132</i>	0.922	0.850 (corrected R ² = 0.847)	3.14 × 10 ⁻⁶⁷	6.49
Blood* + Bones autopsies + Teeth + Buccal swabs	Group 3	Sanger sequencing	<i>ELOVL2</i> CpG3, <i>ELOVL2</i> CpG5, <i>ELOVL2</i> CpG8, <i>ELOVL2</i> CpG9	0.778	0.605 (corrected R ² = 0.598)	4.26 × 10 ⁻⁴³	12.09
	Group 2	SNaPshot	<i>ELOVL2</i> , <i>KLF14</i>	0.870	0.756 (corrected R ² = 0.754)	1.95 × 10 ⁻⁶¹	9.02
Blood** + Bones autopsies + Teeth	Group 2	Sanger sequencing	<i>C1orf132</i> CpG4, <i>FHL2</i> CpG5, <i>FHL2</i> CpG6, <i>FHL2</i> CpG8, <i>ELOVL2</i> CpG5	0.797	0.635 (corrected R ² = 0.620)	2.71 × 10 ⁻²⁴	7.27

*Blood samples from living and deceased individuals; **Blood samples from deceased individuals.

In forensic contexts, the developed APMs can be potentially used in several situations such as: i) cases of migration, in which the age of the minor is unknown; ii) cases in which only human remains are present, frequently associated with mass disaster or crimes against humanity, including war crimes and human rights violation; or iii) cases of unidentified fresh bodies.

In cases of migration, the developed models for buccal swabs (a less invasive collection of the biological sample) or the APMs developed using blood samples from living individuals, can be useful. The APMs developed using blood from living individuals allowed a high accuracy in age predictions, particularly the APM developed with methylation data captured by Sanger sequencing that included *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 (MAD = 5.35 years) and the APM developed using SNaPshot methodology, with methylation information of *ELOVL2*, *FHL2* and *C1orf132* genes (MAD = 4.25 years). While, for buccal swabs the developed APMs using Sanger sequencing, with two CpGs from *ELOVL2* gene (MAD = 8.32 years), and using the SNaPshot methodology, with the CpG located at *TRIM59* gene (MAD = 6.73 years), showed lower accuracy as less CpGs or genes were selected.

For cases in which only human remains are present, the tissue-specific models developed for bones and teeth can be promising in age estimations. Bones and teeth represent one of the most important evidences of DNA preserved in forensic contexts, being sometimes the last evidence of the individuals. Among the developed models, the APM built for bones collected during autopsies through Sanger sequencing methodology, that includes six CpGs located at *ELOVL2*, *C1orf132* and *EDARADD* genes (*ELOVL2* CpG5, *ELOVL2* CpG7, *ELOVL2* CpG6, *C1orf132* CpG1, *EDARADD* CpG3, *EDARADD* CpG4), was the most promising, obtaining a high accuracy of 2.56 years. Anthropological age estimation methods for adults based in skeletal age cannot overcome this value of

accuracy and in cases of missing bones, or when degraded bones are present anthropological methods cannot be applied. For tooth samples, the APM developed with methylation data of only two CpGs, located at *ELOVL2* and *KLF14* genes captured by SNaPshot methodology, allowed to estimate age with a moderate accuracy of 7.07 years. This model can be promising for age estimations in adults and older adults, but cannot overcome the age prediction made by odontological approaches in children.

In addition, if we were dealing with fresh bodies, sometimes without any kind of documentation or possibly any complaints (as cases of disasters with refugees), the APMs developed using blood samples from deceased individuals can be advantageous, helping the identification. In these cases, identification is a paramount issue related to humanitarian, civil and legal aspects. Age estimation can direct the experts and lead to a positive identification.

Even more, we can hypothesize that our developed APM for blood samples from living individuals can be applied to fresh bloodstains in forensic contexts due to DNAm stability, as it was demonstrated for developed models in previous studies (Zbieć-Piekarska *et al.*, 2015a; Huang *et al.*, 2015; Thong *et al.*, 2017). Huang *et al.* (2015) observed no statistically significant differences in age prediction between blood samples and bloodstains, as well as in the comparison of predicted age on the basis of different period bloodstains. These results were similar to the observed by Zbieć-Piekarska *et al.* (2015a) in which the rate of correct predictions in bloodstains seems to be not related with the time of storage. Also, Thong *et al.* (2017) showed that the MAD values remain similar when applying the blood model to bloodstains. Based on this, it is also possible that our models developed for blood samples from living individuals could be applied to forensic casework in cases of bloodstains.

Lastly, also the multi-tissue APMs can be advantageous in forensic cases in which the origin of the sample is unknown, since their application do not require prior knowledge of the tissue type in question. Our developed multi-tissue APMs revealed a moderate accuracy of 6 to 7 years combining blood from living and deceased individuals, bone and tooth samples (group 1 either in Sanger or in SNaPshot).

In the present study, the *ELOVL2* gene revealed to be the strongest age-correlated gene in almost all the tissues investigated. DNAm levels captured by Sanger sequencing revealed for the nine CpGs located at *ELOVL2* the best performance of the change in DNAm with aging in blood samples from living individuals ($0.850 \leq R \leq 0.936$), blood samples from deceased individuals ($0.663 \leq R \leq 0.785$), buccal swabs ($0.485 \leq R \leq 0.823$) and bones from autopsies ($0.590 \leq R \leq 0.852$). DNAm levels captured in teeth and bones from BDS seem to be an exception because *ELOVL2* revealed no significant p-values in almost all CpGs. Moreover, in the tissue-specific APMs developed by Sanger, we observed that CpGs from *ELOVL2* are always included in the multi-locus models developed using blood samples (living and deceased individuals) and bones from autopsies (**Table 5.3**). Also, in the tissue-specific models developed by SNaPshot we can observe that the *ELOVL2* is always present in the multi-locus models developed for blood samples from living and deceased individuals and tooth samples (**Table 5.3**). Concerning the multi-tissue models, developed through Sanger sequencing or SNaPshot methodologies, all the multi-locus APMs included CpGs located at *ELOVL2* (**Table 5.4**). Also in all groups of combined samples, simple linear regression models revealed the highest age-correlation value for *ELOVL2* CpG5 (Group 1: $R = 0.706$; Group 2: $R = 0.563$; Group 3: $R = 0.736$) using Sanger sequencing. In groups of combined samples using SNaPshot data, the CpG located at *ELOVL2* revealed the higher age-correlation

value (Group 1: $R = 0.779$; Group 2: $R = 0.793$). Hence, in accordance with the previous studies, the *ELOVL2* gene has shown to be the most promising marker for age assessment. It has been demonstrated that this gene exhibits reliable age-associated changes in different tissues such as blood, teeth, saliva or bone (Bekaert *et al.*, 2015; Naue *et al.*, 2018; Jung *et al.*, 2019; Gopalan *et al.*, 2019).

Until now, few studies have focused in blood samples from deceased individuals (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Naue *et al.*, 2018; Pfeifer *et al.*, 2020). Moreover, when Bekaert *et al.* (2015a), Hamano *et al.* (2016) and Pfeifer *et al.* (2020) compared DNAm data obtained from living and deceased individuals, no significant differences were observed. However, our study revealed that potential differences in methylation status between blood samples from living and deceased individuals could exist, since the highest age-correlated CpGs were different for some genes between the two groups when assessed through the bisulfite sequencing methodology. Moreover, the correlation between DNAm and age obtained in genes *ELOVL2*, *EDARADD*, *FHL2* and *PDE4C* is lower in blood samples of deceased individuals *versus* living individuals. Also, through the SNaPshot assay, DNAm levels of *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes showed higher age correlation values in blood samples from living individuals ($0.79 \leq R \leq 0.95$) compared to blood samples from deceased individuals ($0.57 \leq R \leq 0.79$). Hence, our study suggests the possible influence of *postmortem* changes that can alter the methylation status among specific loci, although this issue has not yet been clarified until now.

Moreover, our study showed that the APMs developed in blood samples from living individuals cannot be applied to deceased individuals with a similar accuracy (Chapter 4. Results and discussion: A. DNA methylation age estimation in blood samples,

3.4. *Applicability of the developed APMs for blood samples from living and deceased individuals*). Indeed, when we applied the final APM built with methylation information captured by Sanger sequencing from living individuals (MAD = 5.35 years) to the independent set of 51 blood samples from deceased individuals, we obtained a higher value of MAD (9.72 years), which represents a decrease of the model accuracy in this set of samples. In the same way, applying the final multi-locus APM developed for living individuals using the SNaPshot methodology (MAD = 4.25 years) to the methylation information captured in the set of deceased individuals, a MAD of 7.84 years was observed. Contrary, it seems that the APM developed for blood samples from deceased individuals with four genes by Sanger sequencing with a MAD of 6.42 years, when applied to the training set of living individuals demonstrated a similar high accuracy (MAD = 6.10 years). In concordance, the model developed for blood samples from deceased individuals by SNaPshot methodology (MAD = 5.36 years) when applied to the set of living individuals revealed a similar MAD of 5.40 years. In addition, when the life and death status of the donor is not known, it is more advantageous to use the APMs developed in deceased individuals or developed for a combined training set of blood samples from living and deceased individuals, as the models developed in *Chapter 4*, with MAD of 4.97 years and 6.21 years for SNaPshot and Sanger sequencing, respectively.

Regarding to DNAm changes with the increase of age, our study showed in blood samples from living and deceased individuals and also in buccal swabs that the prediction accuracy depends on the chronological age of individuals. Higher MAD values and lower percentage of correct predictions were obtained in older ages, in concordance with previous studies (Bekaert *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015a, 2015b; Hamano *et al.*, 2016; Pfeifer *et al.*, 2020). This can be explained by inter-individual differences in

the rate of methylation changes during aging due to environmental, diseases and stochastic factors being these differences slight in youths and accumulating with age (Jaenisch and Bird, 2003; Boks *et al.*, 2009; Heyn *et al.*, 2012; Spólnicka *et al.*, 2017).

This enables that the APMs are more accurate in younger than in older individuals (Bekaert *et al.*, 2015a; Hamano *et al.*, 2016; Pfeifer *et al.*, 2020).

Another relevant issue in this field is the possible influence of sex in DNAm levels of the age-correlated markers. To date, there is no consensus for a relationship between age-associated DNAm levels and sex (Bekaert *et al.*, 2015a, 2015b; Zbieć-Piekarska *et al.*, 2015b; Huang *et al.*, 2015; Freire-Aradas *et al.*, 2018; Márquez-Ruiz *et al.*, 2020). In our study, very slight differences or no significant differences between males and females were obtained for DNAm levels of age-related CpGs located at *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C*, *C1orf132*, *TRIM59* and *KLF14* genes from blood, teeth, bones and buccal swabs. This result is in concordance with previous studies in these or others genes (Bekaert *et al.*, 2015a; Huang *et al.*, 2015; Zbieć-Piekarska *et al.*, 2015b; Freire-Aradas *et al.*, 2018; Daunay *et al.*, 2019).

Regarding to population specific changes in DNAm, our study using the SNaPshot methodology suggests some differences in DNAm levels between blood samples of Portuguese individuals and other population groups, as Korean and Polish individuals (Zbieć-Piekarska *et al.*, 2015b; Cho *et al.*, 2017; Jung *et al.*, 2019). Such differences suggest that specific markers can be more adequate to different population groups to explain DNAm age-related variation. This population specificity of DNAm levels can justify the usefulness of replication and validation studies of proposed age-related markers as well as genotyping methods in different populations before forensic applications.

We should note that our study suffers from some limitations that should be highlighted:

i) The major drawback was the limited number of collected samples, mainly in bones and teeth. On regards of swab samples, although easier to obtain, the extracted amount of DNA was often insufficient for age-related DNAm analysis. Larger sample sets have greater statistical power and may be more representative of DNAm changes related to different age groups and different types of tissues. Hence, possibly for all the training sets used in the present research, larger sample sizes with larger age ranges could be helpful in obtaining more accurate APMs. However, we should note that other studies in the forensic field used similar sample sizes in the training set, including: Huang *et al.* (2015) with 89 blood samples (aged 9-75 years old); Alghanim *et al.* (2017) with 72 blood samples (aged 5-73 years old); Naue *et al.* (2018) with a set of 29 deceased individuals (including blood, bone and other tissues, aged 0-87 years old); Giuliani *et al.* (2016) with 21 modern teeth (aged 17-77 years old). In any case, in regards to our study, a number of difficulties in sample collection should be noted: generally, it is easier to collect samples from living individuals, including blood, buccal swabs, or teeth (except for living children in which there was great difficulty in obtaining tooth samples). When dealing with deceased individuals, the difficulty of obtaining samples (blood, bones and tooth samples) should be considered because samples of young individuals are rarer and only available in cases of road accident, for example. Moreover, ethical and bureaucratic issues can lead to a delay in collecting samples for all age ranges, as only individuals who do not declare in RENNDA their life-long opposition can be used for scientific researches.

ii) Other factors, such as the existence of some diseases or clinical conditions or even some routines such as smoking or drinking, may interfere with methylation data. In

samples from deceased individuals, despite having access to medical reports of each case, information related to possible clinical conditions was unknown in many cases.

iii) Samples collected from the CEI/XXI suffer also from some limitations as the fact of skeletons do not have many teeth due to advanced age. Consequently, no tooth samples have been collected for this study. Moreover, skeletons may have been exposed to several *postmortem* conditions leading to DNA degradation and loss. This may explain the small amount of DNA extracted from skeletonized individuals in the present study.

iv) Bones collected from BDS can be affected by the embalming method. In this work, the bone fragment was collected after the embalming method with Thiel (Eisma *et al.*, 2013). This can possibly influence DNAm analysis. Moreover, bone samples from BDS have different PMI, which could be a factor of DNAm changes. However, in our study, we observed consistent DNAm changes between individuals from BDS with the same or similar chronological age but different PMIs.

v) The use of different methodologies for evaluation of DNAm levels across the studies can influence the accuracy of APMs. In particular, bisulfite sequencing or the SNaPshot methodologies are semi-quantitative methods and thus may not be the optimal tool for precise DNAm analysis.

Chapter 6. Conclusions and future directions

Chapter 6. Conclusions and future directions

This study, which is based on the assessment of DNAm levels of highly age-correlated genes, can have a significant contribution to the knowledge of forensic age estimation. Several types of samples from Portuguese individuals were examined allowing the selection of the best age-correlated markers in each tissue type. Consequently, we developed several tissue-specific APMs potentially useful for several forensic contexts as in cases of migration, cases associated with mass disaster, crimes against humanity or cases of unidentified fresh bodies. Moreover, combining methylation information of different biological tissues, several multi-tissue APMs useful in cases in which the origin of the sample is unknown have been built. Our study revealed that the bone-APMs based on DNAm levels can be more advantageous for forensic age estimation than the traditional anthropological methods, currently used for adults age assessment, since a higher age prediction accuracy has been obtained (between 2.56 years to 7.18 years in bone samples from autopsies). However, in future studies, the validation of these APMs built using bone samples should be made in a larger sample set.

General conclusions from this study are reported below:

- ❖ Our results are in accordance with the previous studies showing that DNAm levels at specific CpG sites can be used to predict the age of an individual. The DNAm age or epigenetic age was found strongly associated with chronological age and this may become a relevant tool that can help in forensic positive identification.
- ❖ In accordance with previous reports, our study revealed that the age prediction accuracy depends on the chronological age of individuals, being observed higher

MAD values and a lower percentage of correct predictions in older age groups in comparison with younger age groups.

- ❖ In general, similar DNAm patterns were observed using both Sanger sequencing and SNaPshot methodologies, suggesting that both methods can be used in forensic casework.
- ❖ Our data showed that DNAm levels could be slightly influenced by population-specific differences, at least in blood samples. In another way, it was not observed significant DNAm sex differences in each tissue type under study.
- ❖ DNAm levels captured in bone samples from BDS do not seem to be influenced by the PMI.
- ❖ It was observed that *postmortem* alterations in blood samples can influence DNAm levels of the age-correlated markers and consequently the accuracy of APMs. This can be a challenge for forensic purposes when applying APMs developed in the blood samples from living individuals to blood from deceased individuals, leading to misinterpretation of age predictions. However, it has been demonstrated that APMs developed for blood samples from deceased individuals can be accurately applied to blood from living individuals.
- ❖ Generally, DNAm levels in blood samples from living individuals captured by both Sanger sequencing and SNaPshot methodologies show higher age correlation values compared to the other tissue types, including blood samples from deceased individuals.
- ❖ The tissue-specific features of DNAm led to different selection of genes and/or CpGs according to the tissue under analysis; consequently, specific APMs should be developed for each tissue type being more accurate for age predictions.

- ❖ In accordance with previous studies, the number of predictor variables included in tissue-specific and multi-tissue APMs influences the accuracy of age predictions. However, our study revealed that the methodology used for assessment of DNAm levels can also influence the prediction accuracy in each tissue type. The best tissue-specific APMs developed in blood samples (living and deceased individuals) have been built by SNaPShot assay with a lower number of variables comparing with the models developed by Sanger sequencing. However, the best model developed for bones from autopsies was built using Sanger sequencing and a larger number of predictor variables.
- ❖ The identification of universal markers for age prediction, showing similar patterns of DNAm across different tissues, remains a challenge in the forensic field. Meanwhile, our multi-tissue APMs with a combination of three to nine CpGs developed for blood, tooth and bone samples can achieve a very strong age correlation value and an accuracy of 6 to 7 years.
- ❖ In general, the best age marker is the *ELOVL2* gene, being selected for almost all the developed models (tissue-specific APMs and multi-tissue APMs) and showing similar high degree of association with age in the different biological samples.

In regards of future studies to best complement and validate those results obtained in the present work, it would be important:

- to validate and test larger sample sets, mainly for bones and teeth, with the inclusion of all age ranges to improve the reliability of the tissue-specific epigenetic clocks;
- to validate the developed multi-tissue epigenetic clocks in larger independent sample sets;

- to test the accuracy of our tissue-specific and multi-tissue APMs in other ancestry groups;

- to explore the accuracy of the developed epigenetic clocks in individuals with some known diseases or to build specific clocks for certain diseases;

- to test other methodologies for DNAm assessment as massively parallel sequencing (MPS), which have shown successful results with limited amount and/or highly degraded DNA. The use of MPS for DNAm analysis on bone samples from skeletonized individuals can be a more reliable tool, allowing the successful development of DNAm-APMs.

Moreover, to give new insights in the field of DNAm research for age estimation, a number of different suggestions can be addressed in terms of future research:

- to build guidelines for implementation of DNAm age models in forensic laboratories considering ethical and bureaucratic issues, and technical and methodological aspects related to correct DNAm evaluation, as suggested by Bell *et al.* (2019). To the best of our knowledge, despite the increasing development of DNAm age research, it is not yet clarified how these models can be used in routine forensic casework;

-to test additional epigenetic features as chromatin modifications associated with the aging process or to test the combination of several approaches as anthropological, chemical, and epigenetic approaches in order to join many age indicators for the development of more accurate APMs. Aging being a complex biological process is difficult to predict. Age estimation can be improved by the use of a multidisciplinary approach.

Chapter 7. Bibliography

Chapter 7. Bibliography

A

Adserias-Garriga J, Thomas C, Ubelaker DH, Zapico SC. When forensic odontology met biochemistry: Multidisciplinary approach in forensic human identification. *Arch Oral Biol.* 2018;87:7–14.

Adserias-Garriga J, Zapico SC. Age Assessment in Forensic Cases: Anthropological, Odontological and Biochemical Methods for Age Estimation in the Dead. *Mathews J Foren.* 2018;1(1):001.

Adserias-Garriga J. Age estimation: A multidisciplinary approach. Elsevier; 2019a.

Adserias-Garriga J, Wilson-Taylor R. Skeletal age estimation in adults. In: Adserias-Garriga J, editor. Age estimation: A multidisciplinary approach. Elsevier; 2019. pp. 55-68.

Adserias-Garriga J. Odontological approach of age estimation. In: Adserias-Garriga J, editor. Age estimation: A multidisciplinary approach. Elsevier; 2019b. pp. 77-84.

Albert AM, Wright CL. DNA prediction in Forensic Anthropology and the Identity Sciences. *G J Anthropol Res.* 2015;2:1-6.

Alghanim H, Antunes J, Silva DSBS, Alho CS, Balamurugan K, McCord B. Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. *Forensic Sci Int Genet.* 2017;31:81-88. DOI: 10.1016/j.fsigen.2017.07.011.

Aliferi A, Ballard D, Gallidabino MD, Thurtle H, Barron L, Syndercombe Court D. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Sci Int Genet.* 2018;37:215-226. <https://doi.org/10.1016/j.fsigen.2018.09.003>.

AlQahtani SJ, Hector MP, Liversidge HM. Brief communication: the London atlas of human tooth development and eruption. *Am J Phys Anthropol.* 2010;142(3): 481–90.

Alsaleh H, McCallum NA, Halligan DL, Hadrill PR. A multi-tissue age prediction model based on DNA methylation analysis. *Forensic Sci Int Genet Suppl Ser.* 2017;6:e62-e64. <https://doi.org/10.1016/j.fsigss.2017.09.056>.

Page | 324 Ambatipudi S, Horvath S, Perrier F, Cuenin C, Hernandez-Vargas H, Le Calvez- Kelm F, Durand G, Byrnes G, Ferrari P, Bouaoun L, Sklias A, Chajes V, Overvad K, Severi G, Baglietto L, Clavel-Chapelon F, Kaaks R, Barrdahl M, Boeing H, Trichopoulou A, Lagiou P, Naska A, Masala G, Agnoli C, Polidoro S, Tumino R, Panico S, Dolle M, Peeters PHM, Onland-Moret NC, Sandanger TM, Nost TH, Weiderpass E, Quiros JR, Agudo A, Rodriguez-Barranco M, Huerta Castano JM, Barricarte A, Fernandez AM, Travis RC, Vineis P, Muller DC, Riboli E, Gunter M, Romieu I, Herceg Z. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer.* 2017;75:299–307.

Amory S, Huel R, Bilić A, Loreille O, Parsons TJ. Automatable full demineralization DNA extraction procedure from degraded skeletal remains. *Forensic Sci Int Genet.* 2012;6:398e406.

Applied Biosystems. Manual: Using the SNaPshot® Multiplex System with the POP-7™ Polymer on Applied Biosystems 3730/3730xl DNA Analyzers and 3130/3130 (English). 2005. Available from: <https://www.thermofisher.com/order/catalog/product/4323159#/4323159>.

Applied Biosystems. Quick Reference Card: ABI PRISM® SNaPshot™ Multiplex System. 2010. Available from: <https://www.thermofisher.com/order/catalog/product/4323159#/4323159>.

Azevedo J, Neves A, Cunha, E. Estimativa da idade em indivíduos vivos indocumentados. Norma Procedimental INMLCF 018. 2019.

B

Bacalini MG, Deelen J, Pirazzini C, De Cecco M, Giuliani C, Lanzarini C, Ravaioli F, Marasco E, van Heemst D, Suchiman HED, Slieker R, Giampieri E, Recchioni R, Marcheselli F, Salvio S, Vitale G, Olivieri F, Spijkerman AMW, Dollé MET, Sedivy JM, Castellani G, Franceschi C, Slagboom PE, Garagnani P. Systemic Age-Associated DNA Hypermethylation of ELOVL2 Gene: In Vivo and In Vitro Evidences of a Cell Replication Process. *J Gerontol A Biol Sci Med Sci.* 2017;72(8):1015-1023. doi: 10.1093/gerona/glw185.

Baccino E, Cunha E, Cattaneo C. Aging the dead and the living. Elsevier. 2013;42-48.

Baccino E, Ubelaker DH, Hayek L-AC, Zerilli A. Evaluation of Seven Methods of Estimating Age at Death from Mature Human Skeletal Remains. *J Forensic Sci.* 1999;931-936.

Becker J, Mahlke NS, Reckert A, Eickhoff SB, Ritz-Timme S. Age estimation based on different molecular clocks in several tissues and a multivariate approach: an explorative study. *Int J Legal Med.* 2019;1-13. Page | 325

Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, Christensen BC, Gladyshev VN, Heijmans BT, Horvath S, Ideker T, Issa JJ, Kelsey KT, Marioni RE, Reik W, Relton CL, Schalkwyk LC, Teschendorff AE, Wagner W, Zhang K, Rakyan VK. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* 2019;20(1):249. doi: 10.1186/s13059-019-1824-y.

Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics.* 2015a;10:922–930.

Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples. *Forensic Sci Int Genet Suppl Ser* 2015b;5:e144-e145.

Bergsma T, Rogaeva E. DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. *Neurosci Insights.* 2020;21(15):2633105520942221.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol.* 2010;28:1045–1048.

Blau S. How traumatic: a review of the role of the forensic anthropologist in the examination and interpretation of skeletal trauma. *Aust J Forensic Sci.* 2016;49(3):261-280. DOI: 10.1080/00450618.2016.1153715.

Blau S. It's all about the context: reflections on the changing role of forensic anthropology in medico-legal death investigations. *Aust J Forensic Sci.* 2018. DOI: 10.1080/00450618.2017.1422022.

Bocklandt S, Lin W, Sehl ME, Sanchez FJ, Sinsheimer JS, Horvath S, Vilain E. Epigenetic predictor of age. *PLoS One*. 2011;6(6):1-6.

Brooks S, Suchey JM. Skeletal age determination based on the os pubis: A comparison of the Acsádi-Nemeskéri and Suchey-Brooks methods. *Hum Evol*. 1990;5:227–238.

Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, Khan RS, Ophoff RA. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*. 2009;4.

Botelho S. Assessing ancestry through genetics: an attempt using skeletonized human remains. 2020.

Breitling LP, Saum KU, Perna L, Schottker B, Holleczeck B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. *Clin Epigenetics*. 2016;8:21. doi: 10.1186/s13148-016-0186-5.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Paabo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. 2010;38: e87.

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res*. 2007;35(17):5717-28. doi: 10.1093/nar/gkm588.

Buckberry JL, Chamberlain AT. Age estimation from the auricular surface of the ilium: a revised method. *Am J Phys Anthropol*. 2002;119(3):231-239.

Butler MG, Tilburt J, DeVries A, Muralidhar B, Aue G, Hedges L, Atkinson J, Schwartz H. Comparison of chromosome telomere integrity in multiple tissues from subjects at different ages. *Cancer Genet Cytogenet*. 1998;105(2):138-44. doi: 10.1016/s0165-4608(98)00029-6.

C

Cabo LL. DNA analysis and the classic goal of Forensic Anthropology. In: Dikmaat DC, editor. *A Companion to Forensic Anthropology* Blackwell Publishing LTD; 2012. pp. 447-461.

Cameriere R, Ferrante L, Belcastro MG, Bonfiglioli B, Rastelli E, Cingolani M. Age estimation by pulp/tooth ratio in canines by peri-apical X-rays. *J Forensic Sci.* 2007;52(1):166-170.

Cameriere R, De Luca S, Alemán I, Ferrante L, Cingolani M. Age estimation by pulp/tooth ratio in lower premolars by orthopantomography. *Forensic Sci Int.* 2012;214(1-3):105-112.

Carracedo A. Forensic Genetics: History. In: Houck MM, editor. *Forensic Biology.* Elsevier Ltd; 2013. pp. 19-22.

Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. *Nature.* 2019;571:489-499.

Cattaneo C, De Angelis D, Ruspa M, Gibelli D, Cameriere R, Grandi M. How old am I? Age estimation in living adults: a case report. *J Forensic Odontostomatol.* 2008;27(2):39-43.

Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, Roetker NS, Just AC, Demerath EW, Guan W, Bressler J, Fornage M, Studenski S, Vandiver AR, Moore AZ, Tanaka T, Kiel DP, Liang L, Vokonas P, Schwartz J, Lunetta KL, Murabito JM, Bandinelli S, Hernandez DG, Melzer D, Nalls M, Pilling LC, Price TR, Singleton AB, Gieger C, Holle R, Kretschmer A, Kronenberg F, Kunze S, Linseisen J, Meisinger C, Rathmann W, Waldenberger M, Visscher PM, Shah S, Wray NR, McRae AF, Franco OH, Hofman A, Uitterlinden AG, Absher D, Assimes T, Levine ME, Lu AT, Tsao PS, Hou L, Manson JE, Carty CL, LaCroix AZ, Reiner AP, Spector TD, Feinberg AP, Levy D, Baccarelli A, van Meurs J, Bell JT, Peters A, Deary IJ, Pankow JS, Ferrucci L, Horvath S. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY).* 2016;8(9):1844-1865. doi: 10.18632/aging.101020.

Cho S, Ge J, Seo SB, Kim K, Lee HY, Lee SD. Age estimation via quantification of signal-joint T cell receptor excision circles in Koreans. *Leg Med.* 2014;16(3):135-8. doi: 10.1016/j.legalmed.2014.01.009.

Cho S, Jung SE, Hong SR, Lee EH, Lee JH, Lee SD, Lee HY. Independent validation of DNA-based approaches for age prediction in blood. *Forensic Sci Int Genet.* 2017;29:250-256. doi: 10.1016/j.fsigen.2017.04.020.

Ciccarone F, Tagliatesta S, Caiafa P, Zampieri M. DNA methylation dynamics in aging: how far are we from understanding the mechanisms? *Mech Ageing Dev.* 2018;174:3-17.

Conway K, Edmiston SN, Parrish E, Bryant C, Tse CK, Swift-Scanlan T, McCullough LE, Kuan PF. Breast tumor DNA methylation patterns associated with smoking in the Carolina Breast Cancer Study. *Breast Cancer Res Treat.* 2017;163(2):349-361. doi: 10.1007/s10549-017-4178-8.

Cortopassi GA, Shibata D, Soong N-W, Arnheim N. A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. *Proc Natl Acad Sci USA* 1992;89:7370-7374.

Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, Lynnerup N, Cattaneo C. The problem of aging human remains and living individuals: a review. *Forensic Sci Int.* 2009;193:1-13. <https://doi.org/10.1016/j.forsciint.2009.09.008>.

Cunha E. A Antropologia Forense passo a passo. In A. Gomes. A enfermagem forense. Lidel 2012;280-288.

Cunha E, Ferreira T. Antropologia Forense. In: Corte-Real F e Cunha E, editors. *Tratado de Medicina Legal*. In press. 2021.

Cunha E. Devolvendo a identidade: A Antropologia Forense no Brasil. *Cienc Cult.* 2019;71(2):30-34. <http://dx.doi.org/10.21800/2317-66602019000200011>.

Cunningham CA, Scheuer L, Black SM. *Developmental juvenile osteology*. 2nd ed. London, Elsevier; 2016.

Cunningham CA. Skeletal age estimation in juveniles and subadults. In: Adserias-Garriga J, editor. *Age estimation: A multidisciplinary approach*. Elsevier; 2019. pp. 41-51.

D

Daunay A, Baudrin LG, Deleuze JF, How-Kit A. Evaluation of six blood-based age prediction models using DNA methylation analysis by pyrosequencing. *Sci Rep.* 2019;9(1):8862. doi: 10.1038/s41598-019-45197-w.

de Boer HH, Blau S, Delabarde T, Hackman L. The role of forensic anthropology in disaster victim identification (DVI): recent developments and future prospects. *Forensic Sci Res.* 2019;4(4): 303-315, DOI: 10.1080/20961790.2018.1480460.

de Boer HH, Obertová Z, Cunha E, Adalian P, Baccino E, Fracasso T, Kranioti E, Lefèvre P, Lynnerup N, Petaros A, Ross A, Steyn M, Cattaneo C. Strengthening the role of forensic anthropology in personal identification: Position statement by the Board of the Forensic Anthropology Society of Europe (FASE). *Forensic Sci Int.* 2020;315:110456. doi: 10.1016/j.forsciint.2020.110456. Page | 329

Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Develop.* 2011;25:1010–1022.

Deans C, Maggert KA. What do you mean, "epigenetic"? *Genetics.* 2015;199(4):887-96. doi: 10.1534/genetics.114.173492.

Decreto-Lei n.º 274/99, de 22 de Julho. Available from: <https://dre.pt/application/file/a/354775>.

Diba LZ, Ardebili SMM, Jalal Gharesouran J, Houshmand M. Age-related decrease in mtDNA content as a consequence of mtDNA 4977bp deletion. *Mitochondrial DNA A DNA Mapp Seq Anal.* 2015;27(4):3008-12. doi: 10.3109/19401736.2015.1063046.

Dugué P-A, Bassett JK, Joo JE, Jung C-H, Wong EM, Margarita Moreno-Betancur M, Schmidt D, Makalic E, Li S, Severi G, Hodge AM, Buchanan DD, English DR, Hopper JL, Southey MC, Giles GG and Milne RL. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int J Cancer.* 2018;142:1611–1619.

E

Eipel M, Mayer F, Arent T, Ferreira MRP, Birkhofer C, Gerstenmaier U, Costa IG, Ritz-Timme S, Wagner W. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. *Aging.* 2016;8(5):1034-48.

Eisma R, Lamb C, and Soames RW. From formalin to Thiel embalming: What changes? One anatomy department's experiences. *Clin Anat.* 2013;26(5):564–571.

Elamin F, Liversidge HM. Malnutrition Has No Effect on the Timing of Human Tooth Formation. *PLoS One*. 2013;8(8):e72274.

F

Page | 330 Ferreira MT, Vicente R, Navega D, Gonçalves D, Curate F, Cunha E. A new forensic collection housed at the University of Coimbra, Portugal: The 21st century identified skeletal collection. *Forensic Sci Int*. 2014;245:202.e1-5. doi: 10.1016/j.forsciint.2014.09.021.

Ferreira MT, Coelho C, Makhoul C, Navega D, Gonçalves D, Cunha E, Curate F. New data about the 21st Century Identified Skeletal Collection (University of Coimbra, Portugal). *Int J Legal Med*. 2020. doi: 10.1007/s00414-020-02399-6.

Field AE, Robertson NA, Wang T, Havas A, Ideker T, and Adams PD. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell*. 2018;71:882-895.

Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, Colicino E, Dugué PA, Artaud F, McKay GJ, Jeong A, Mishra PP, Nøst TH, Krogh V, Panico S, Sacerdote C, Tumino R, Palli D, Matullo G, Guarrera S, Gandini M, Bochud M, Dermitzakis E, Muka T, Schwartz J, Vokonas PS, Just A, Hodge AM, Giles GG, Southey MC, Hurme MA, Young I, McKnight AJ, Kunze S, Waldenberger M, Peters A, Schwettmann L, Lund E, Baccarelli A, Milne RL, Kenny RA, Elbaz A, Brenner H, Kee F, Voortman T, Probst-Hensch N, Lehtimäki T, Elliot P, Stringhini S, Vineis P, Polidoro S; BIOS Consortium; Lifepath consortium. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. *Aging (Albany NY)* 2019;11(7):2045-2070. doi: 10.18632/aging.101900.

Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum Mol Genet*. 2014;23 (5):1186-1201.

Fleckhaus J, Freire-Aradas A, Rothschild MA, Schneide PM. Impact of genetic ancestry on chronological age prediction using DNA methylation analysis. *Forensic Sci Int Genet Suppl Ser*. 2017;6: e399–e400.

Franklin D. Forensic age estimation in human skeletal remains: current concepts and future directions. *Leg Med*. 2010;12:1–7.

Franklin D, Flavel A, Noble J, Swift L, Karkhanis S. Forensic age estimation in living individuals: methodological considerations in the context of medico-legal practice. *Research and Reports in Forensic Medical Science* 2015;5:53-66.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Page | 331
Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci USA*. 2005;102:10604–10609.

Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. *Genome Biol*. 2012;13(2):R8:2-12.

Freire-Aradas A, Phillips C, Mosquera-Miguel A, Girón-Santamaría L, Gómez-Tato A, Casares de Cal M, Álvarez-Dios J, Ansedo-Bermejo J, Torres-Español M, Schneider PM, Pośpiech E, Branicki W, Carracedo Á, Lareu MV. Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci Int Genet*. 2016;24:65-74.

Freire-Aradas A, Phillips C, Lareu MV. Forensic individual age estimation with DNA: from initial approaches to methylation tests. *Forensic Sci Rev*. 2017;29:121–144.

Freire-Aradas A, Phillips C, Girón-Santamaría L, Mosquera- Miguel A, Gómez-Tato A, Casares de Cal MÁ, Álvarez-Dios J, Lareu MV. Tracking age-correlated DNA methylation markers in the young. *Forensic Sci Int Genet*. 2018;36:50–59. [https:// doi.org/10.1016/J.FSIGEN.2018.06.011](https://doi.org/10.1016/J.FSIGEN.2018.06.011).

Freire-Aradas A, Pośpiech E, Aliferi A, Girón-Santamaría L, Mosquera-Miguel A, Pisarek A, Ambroa-Conde A, Phillips C, Casares de Cal MA, Gómez-Tato A, Spólnicka M, Woźniak A, Álvarez-Dios J, Ballard D, Court DS, Branicki W, Carracedo Á, Lareu MV. A Comparison of Forensic Age Prediction Models Using Data From Four DNA Methylation Technologies. *Front Genet*. 2020;19(11):932.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*. 1992;89:1827–1831.

G

Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, Di Blasio AMD, Gentilini D, Vitale G, Collino S, Rezzi S, Castellani G, Capri M, Salvioli S, and Franceschi C. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*. 2012;11:1132–1134.

Page | 332

Garvin HM, Stock MK. The Utility of Advanced Imaging in Forensic Anthropology. *Acad Forensic Pathol*. 2016;6(3):499-516. doi: 10.23907/2016.050.

Garvin HM, Passalacqua NV. Current practices by forensic anthropologists in adult skeletal age estimation. *J Forensic Sci*. 2012;57:427–433. doi: 10.1111/j.1556-4029.2011.01979.x.

Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res*. 2008;36(22):e150. doi: 10.1093/nar/gkn691.

Giuliani C, Cilli E, Bacalini MG, Pirazzini C, Sazzini M, Gruppioni G, Franceschi C, Garagnani P, Luiselli D. Inferring chronological age from DNA methylation patterns of human teeth. *Am J Phys Anthropol* 2016;159(4):585-95. doi: 10.1002/ajpa.22921.

Gillespie SL, Hardy LR, Anderson CM. Patterns of DNA methylation as an indicator of biological aging: State of the science and future directions in precision health promotion. *Nurs Outlook*. 2019;67(4):337-344. doi: 10.1016/j.outlook.2019.05.006.

Goodwin W, Linacre A, Hadi S. Biological material-collection, characterization and storage In: *An Introduction to Forensic Genetics*. John Wiley & Sons Ltd; 2007. pp.17-25.

Goel N, Karira P, Garg VK. Role of DNA methylation in human age prediction, *Mech. Ageing Dev*. 2017;166:33–41.

Gopalan S, Gaige J, Henn BM. DNA methylation-based forensic age estimation in human bone. *BioRxiv*. 2019. <https://doi.org/10.1101/801647>.

Gršković B, Zrnec D, Vicković S, Popović M, Mršić G. DNA methylation: the future of crime scene investigation? *Mol Biol Rep*. 2013;40:4349- 4360. doi: 10.1007/s11033-013-2525-3.

Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res*. 2001;29:e65.

H

Hamano Y, Manabe S, Morimoto C, Fujimoto S, Ozeki M, Tamaki K. Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting. *Leg Med.* 2016;21:5-10.

Hamano Y, Manabe S, Morimoto C, Fujimoto S, Tamaki K. Forensic age prediction for saliva samples using methylation-sensitive high resolution melting: exploratory application for cigarette butts. *Sci Rep.* 2017;7(1):10444. doi: 10.1038/s41598-017-10752-w.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, Klotzle B, Bibikova M, Fan J-B, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.* 2013;49:359–367.

He Y, Ecker JR. Non-CG methylation in the human genome. *Annu Rev Genomics Hum Genet.* 2015;16:55.

Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Javier Carmona F, Puca AA, Sayols S, Pujana MA, Serra-Musach J, Iglesias-Platas I, Formiga F, Fernandez AF, Fraga MF, Heath SC, Valencia A, Gut IG, Wang J, Esteller M. Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci. USA* 2012;109:10522–10527.

Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, Esteller M. DNA methylation contributes to natural human variation. *Genome Res.* 2013;23(9):1363-72. doi: 10.1101/gr.154187.112.

Higgins D, Austin JJ. Teeth as a source of DNA for forensic identification of human remains: a review. *Sci Justice.* 2013;53:433–441.

Hofreiter N, Jaenicke V, Serre D, von Haeseler A, Paabo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* 2001;29:4793–4799.

Hong SR, Jung SE, Lee EH, Shin KJ, Yang WI, Lee HY. DNA methylation based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Forensic Sci Int Genet.* 2017;29:118–125.

Hong SR, Shin K-J, Jung S-E, Lee EH, Lee HY. Platform-independent models for age prediction using DNA methylation data. *Forensic Sci Int Genet.* 2019;38:39–47.

Page | 334 Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14(10):R115.

Horvath S, V. Mah V, Lu AT, Woo JS, Choi OW, Jasinska AJ, Riancho JA, Tung S, Coles NS, Braun J, Vinters HV, Coles LS. The cerebellum ages slowly according to the epigenetic clock. *Aging.* 2015;7:294–306.

Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging (Albany NY).* 2015;7(12):1130-42. doi: 10.18632/aging.100859.

Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet.* 2018; 19:371–384. <https://doi.org/10.1038/s41576-018-0004-3>.

Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, Felton S, Matsuyama M, Lowe D, Kabacik S, Wilson JG, Reiner AP, Maierhofer A, Flunkert J, Aviv A, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Ferrucci L, Matsuyama S, Raj K. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging.* 2018;10(7):1758-1775.

Huang Y, Yan J, Hou J, Fu X, Li L, Hou Y. Developing a DNA methylation assay for human age prediction in blood and bloodstain. *Forensic Sci Int Genet.* 2015;17:129-136 <http://dx.doi.org/10.1016/j.fsigen.2015.05.007>.

I

Illingworth R, Kerr A, Desousa D, Jørgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, Humphray S, Cox T, Langford C, Bird A. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* 2008; 6:e22.

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genetics.* 2010;6(9):e1001134.

Işcan MY, Loth SR, Wright RK. Age estimation from the rib by phase analysis: white males. *J Forensic Sci.* 1984;29(4):1094–1104.

Işcan MY, Loth SR, Wright RK. Metamorphosis at the sternal rib end: a new method to estimate age at death in white males. *Am J Phys Anthropol.* 1984a;65(2):147–156. doi: 10.1002/ajpa.1330650206.

Işcan MY, Loth SR, Wright RK. Age estimation from the rib by phase analysis: white females. *J Forensic Sci.* 1985;30(3):853–863. Page | 335

Işcan MY, Loth SR. Determination of age from the sternal rib in white females: a test of the phase method. *J Forensic Sci.* 1986;31(3):990–999.

J

Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003;33:245–254.

Jakubowska J, Maciejewska A, Pawlowski R. Comparison of three methods of DNA extraction from human bones with different degrees of degradation. *Int J Leg Med.* 2012;126:173e8.

Jiang M, Zhang Y, Fei J, Chang X, Fan W, Qian X, Zhang T, Lu D. Rapid quantification of DNA methylation by measuring relative peak heights in direct bisulfite-PCR sequencing traces. *Lab Invest.* 2010;90(2):282–90. doi: 10.1038/labinvest.2009.132.

Jiang S, Guo Y. Epigenetic Clock: DNA Methylation in Aging. *Stem Cells Int.* 2020;2020:1047896. doi: 10.1155/2020/1047896.

Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. *Aging Cell.* 2015;14:924–932. doi: 10.1111/accel.12349.

Jung SE, Shin KJ, Lee HY. DNA methylation-based age prediction from various tissues and body fluids. *BMB Rep.* 2017;50(11):546–553.

Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci Int Genet.* 2019;38:1–8.

Jylhävä J, Pedersen NL, Hägg S. Biological age predictors. *EBioMedicine* 2017;21:29–36.

K

Kader F, Ghai M. DNA methylation and application in forensic sciences. *Forensic Sci Int.* 2015;249:255–265.

Page | 336 Kananen L, Marttila S, Nevalainen T, Jylhävä J, Mononen N, Kähönen M, Raitakari OT, Lehtimäki T, Hurme M. Aging-associated DNA methylation changes in middle-aged individuals: the Young Finns study. *BMC Genomics.* 2016;17:103.

Karlsson AO, Svensson A, Marklund A, Holmlund G. Estimating human age in forensic samples by analysis of telomere repeats. *Forensic Sci Int Genet Suppl Ser.* 2008;1:569–571. <https://doi.org/10.1016/j.fsigss.2007.10.153>

Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet.* 2015;18:33-48.

Koch MC, Wagner W. Epigenetic-aging-signature to determine age in different tissues. *Aging.* 2011;3:1-10.

Koop BE, Mayer F, Gündüz T, Blum J, Becker J, Schaffrath J, Wagner W, Han Y, Boehme P, Ritz-Timme S. Postmortem age estimation via DNA methylation analysis in buccal swabs from corpses in different stages of decomposition-a "proof of principle" study. *Int J Legal Med.* 2020;135(1):167-173. doi: 10.1007/s00414-020-02360-7.

L

Lacan M, Thèves C, Keyser C, Farrugia A, Baraybar JP, Crubézy E, Ludes B. Detection of age-related duplications in mtDNA from human muscles and bones. *Int J Legal Med.* 2011;125(2):293-300. doi: 10.1007/s00414-010-0440-x.

Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, Suter CM. High-resolution analysis of cytosine methylation in ancient DNA. *PloS one.* 2012;7(1):e30226. doi: 10.1371/journal.pone.0030226.

Lamendin H. Observations on teeth roots in the estimation of age. *Int J Forensic Dent.* 1973;1(1):4–7.

Lamendin H, Baccino E, Humpert JF, Tavernier JC, Nossintchouk RM, Zerilli A. A simple technique for age estimation in adult corpses: the two criteria dental method. *J Forensic Sci.* 1993;37(5):1373–1379.

Lardenoije R, Latrou A, Kenis G, Kompotis K, Steinbusch HWM, Mastroeni D, Coleman P, Lemere CA, Hof PR, van den Hove DLA, Rutten BPF. The epigenetics of aging and neurodegeneration. *Prog Neurobio.* 2015;131:21–64. Page | 337

Larsen F, Gundersen G, Lopez R, Prydz H. CpG Islands as Gene Markers in the Human Genome. *Genomics.* 1992;13:1095-1107.

Lee HC, Pang CY, Hsu HS, Wei YH. Differential accumulations of 4,977 bp deletion in mitochondrial DNA of various tissues in human ageing. *Biochim Biophys Acta.* 1994; 1226:37–43.

Lee HY, Jung S-E, Oh YN, Choi A, Yang WI, Shin K-J. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. *Forensic Sci Int Genet.* 2015;19:28–34.

Lee HY, Lee SD, Shin K-J. Forensic DNA methylation profiling from evidence material for investigative leads. *BMB rep.* 2016;49(7):359-369.

Lee JW, Choung CM, Jung JY, Lee HY, Lim SK. A validation study of DNA methylation-based age prediction using semen in forensic casework samples. *Leg Med.* 2018;31:74-77. <https://doi.org/10.1016/j.legalmed.2018.01.005>.

Lee HY, Hong SR, Lee JE, Hwang IK, Kim NY, Lee JM, Fleckhaus J, Jung SE, Lee YH. Epigenetic age signatures in bones. *Forensic Sci Int Genet.* 2020;46:102261. doi: 10.1016/j.fsigen.2020.102261.

Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, Aviv A, Lohman K, Liu Y, Ferrucci L, Horvath S. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* 2018;10(4):573-591.

Li Y, Tollefsbol TO. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods Mol Biol.* 2011;791:11–21.

Page | 338 Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, Zhang H, Chen Q, Luo R, Chen M, He Y, Jin X, Zhang Q, Yu C, Zhou G, Sun J, Huang Y, Zheng H, Cao H, Zhou X, Guo S, Hu X, Li X, Kristiansen K, Bolund L, Xu J, Wang W, Yang H, Wang J, Li R, Beck S, Wang J, Zhang X. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* 2010;8:e1000533.

Lin Q, Weidner CI, Costa IG, Marioni RE, Ferreira MR, Deary IJ, Wagner W. DNA methylation levels at individual age associated CpG sites can be indicative for life expectancy. *Aging.* 2016;8:394-401. doi: 10.18632/aging.100908.

Liu X, Jiao B, Shen L. The Epigenetics of Alzheimer's Disease: Factors and Therapeutic Implications. *Front Genet.* 2018;9:579. doi:10.3389/fgene.2018.00579.

Liu Z, Leung D, Levine M. Comparative analysis of epigenetic aging clocks from CpG characteristics to functional Associations. *BioRxiv.* 2019;512483.

Loh M, Zhou L, Ng HK, Chambers JC. Epigenetic disturbances in obesity and diabetes: Epidemiological and functional insights. *Mol Metab.* 2019;27:S33-S41.

Love JC, Wiersema JM. Skeletal Trauma: An Anthropological Review. *Acad Forensic Pathol.* 2016;6(3):463-477. doi: 10.23907/2016.047.

Lovejoy CO, Meindl RS, Pryzbeck TR, Mensforth RP. Chronological metamorphosis of the auricular surface of the ilium: a new method for the determination of adult skeletal age death. *Am J Phys Anthropol.* 1985;68(1):15-28.

Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* 2019a;11(2):303-327. doi: 10.18632/aging.101684.

Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, Kooperberg C, Ferrucci L, Hou L, Baccarelli AA, Li Y, Harris SE, Corley J, Taylor A, Deary IJ, Stewart JD, Whitsel EA, Assimes TL, Chen W, Li S, Mangino M, Bell JT, Wilson JG, Aviv A, Marioni RE, Raj K, Horvath S. DNA

methylation-based estimator of telomere length. *Aging* (Albany NY). 2019b;11(16):5895-5923. doi: 10.18632/aging.102173.

Lyko F. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. *Nat Rev Genet.* 2018;19(2):81-92. <https://doi.org/10.1038/nrg.2017.80>.

M

Marano LA, Fridman C. DNA phenotyping: current application in forensic science. *Research and Reports in Forensic Medical Science* 2019;9:1–8.

Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annu Rev Public Health.* 2018;39:309-333. doi: 10.1146/annurev-publhealth-040617-014629.

Martrille L, Ubelaker DH, Cattaneo C, Seguret F, Tremblay M, Baccino E. Comparison of four skeletal methods for the estimation of age at death on white and black adults. *J Forensic Sci.* 2007;52(2):302-7. doi: 10.1111/j.1556-4029.2006.00367.x.

Márquez-Ruiz AB, González-Herrera L, Valenzuela A. Usefulness of telomere length in DNA from human teeth for age estimation. *Int J Legal Med.* 2018;132:353–359. <https://doi.org/10.1007/s00414-017-1595-5>.

Márquez-Ruiz AB, González-Herrera L, Luna JD, Valenzuela A. DNA methylation levels and telomere length in human teeth: usefulness for age estimation. *Int J Legal Med* 2020;134(2):451-459.

Meissner C, Bruse P, Mohamed SA, Schulz A, Warnk H, Storm T, Oehmichen M. The 4977 bp deletion of mitochondrial DNA in human skeletal muscle, heart and different areas of the brain: A useful biomarker or more? *Exp Gerontol.* 2008;43(7):645–652.

Meissner C, Ritz-Timme S. Molecular pathology and age estimation. *Forensic Sci Int.* 2010;203:34–43.

Moore LD, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology.* 2013;38:23–38. <https://doi.org/10.1038/npp.2012.112>.

N

Navega D, Cunha E. Extreme learning machine neural networks for adult skeletal age-at-death estimation. In: Obertová Z, Stewart A, Cattaneo C, editors. *Statistics and Probability in Forensic Anthropology*. Elsevier; 2020. pp. 209-223.

Page | 340

Naue J, Hoefsloot HCJ, Mook ORF, Rijlaarsdam-Hoekstra L, van der Zwalm MCH, Henneman P, Kloosterman AD, Verschure PJ. Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. *Forensic Sci Int Genet*. 2017;31:19-28. doi: 10.1016/j.fsigen.2017.07.015.

Naue J, Sängler T, Hoefsloot HCJ, Lutz-Bonengel S, Kloosterman AD, Verschure PJ. Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. *Forensic Sci Int Genet*. 2018;36:152-159.

Nakhaeizadeh S, Dror IE, Morgan RM. Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. *Sci Justice*. 2014;54(3):208-214

Nwanaji-Enwerem JC, Weisskopf MG, Baccarelli AA. Multi-tissue DNA methylation age: Molecular relationships and perspectives for advancing biomarker utility. *Ageing Res Rev*. 2018;45:15-23.

Nuzzolese E, Di Vella G. Legal background of age estimation for the dead and the living. In: Adserias-Garriga J, editor. *Age estimation: A multidisciplinary approach*. Elsevier; 2019. pp. 17-24.

O

Odzhakov F, Apostolov A. Dental age estimation – Literature review. *J Transl Sci*. 2019;6(3):1-5.

Ohtani S, Yamamoto T. Age estimation by amino acid racemization in human teeth. *J Forensic Sci*. 2010;55(6):1630-1633.

Ou X, Zhao H, Sun H, Yang Z, Xie B, Yanwei S, Wu X. Detection and quantification of the age-related sjTREC decline in human peripheral blood. *Int J Legal Med*. 2011;125: 603–608.

P

Pal S, Tyler JK. Epigenetics and aging. *Sci Adv.* 2016;2(7):e1600584.

Patterson K, Molloy L, Qu W, Clark S. DNA Methylation: Bisulphite Modification and Analysis. *J Vis Exp.* 2011;(56):e3170:1- 9.

Page | 341

Parson W. Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi) Genomics: A Mini-Review. *Gerontology* 2018; doi: 10.1159/000486239.

Parrish RR, Day JJ, Lubin FD. Direct bisulfite sequencing for examination of DNA methylation patterns with gene and nucleotide resolution from brain tissues. *Curr Protoc Neurosci.* 2012;60(1):7.24.1-7.24.12.

Park JL, Kwon OH, Kim JH, Yoo HS, Lee HC, Woo KM, Kim SY, Lee SH, Kim YS. Identification of body fluid-specific DNA methylation markers for use in forensic science. *Forensic Sci Int Genet.* 2014;13:147-53. doi: 10.1016/j.fsigen.2014.07.011.

Park JL, Kim JH, Seo E, Bae DH, Kim SY, Lee HC, Woo KM, Kim YS. Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Sci Int Genet.* 2016;23:64-70.

Pedersen JS, Valen E, Velazquez AM, Parker BJ, Rasmussen M, Lindgreen S, Lilje B, Tobin DJ, Kelly TK, Vang S, Andersson R, Jones PA, Hoover CA, Tikhonov A, Prokhortchouk E, Rubin EM, Sandelin A, Gilbert MT, Krogh A, Willerslev E, Orlando L. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* 2014;24(3):454-66. doi: 10.1101/gr.163592.113.

Pinheiro J, Cunha E, Symes S. Over-interpretation of bone injuries and implications for cause and manner of death. In: Passalacqua NV, Rainwater CW, editors. *Skeletal trauma analysis: Case studies in context* John Wiley & Sons, Ltd; 2015. pp. 27-41.

Pfeifer M, Bajanowski T, Helmus J, Poetsch M. Inter-laboratory adaption of age estimation models by DNA methylation analysis-problems and solutions. *Int J Legal Med.* 2020;134(3):953-961. doi: 10.1007/s00414-020-02263-7.

Prince C, Hammerton G, Taylor AE, Anderson EL, Timpson NJ, Davey Smith G, Munafò MR, Relton CL, Richmond RC. Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence. *Hum Mol Genet.* 2019;28(1):155-165. doi: 10.1093/hmg/ddy316.

Priya E. Methods of skeletal age estimation used by forensic anthropologists in adults: a review. *Forensic Research & Criminology International Journal.* 2017;4(2):41-51.

Q

QIAGEN. Purification of total DNA from compact animal bone using the DNeasy® Blood & Tissue Kit. 2006. Available from: <https://www.google.com/search?client=opera&q=User-Developed+Protocol%3A+Purification+of+total+DNA+from+compact+animal+bone+using+the+DNeasy%26+Tissue+Kit.&sourceid=opera&ie=UTF-8&oe=UTF-8>.

R

Ren F, Li C, Xi H, Wen Y, Huang K. Estimation of human age according to telomere shortening in peripheral blood leukocytes of Tibetan. *Am J Forensic Med Pathol.* 2009;30(3):252-5.

Ritz S, Schlitz HW, Peper C. Postmortem estimation of age at death based on aspartic acid racemization in dentin: its applicability for root dentin. *Int J Legal Med.* 1993(105): 289–293.

Roberti A, Valdes AF, Torrecillas R, Fraga MF, Fernandez AF. Epigenetics in cancer therapy and nanomedicine. *Clin Epigenetics.* 2019;11(1):81. doi: 10.1186/s13148-019-0675-4.

S

Samuel G, Prainsack B. Forensic DNA phenotyping in Europe: views “on the ground” from those who have a professional stake in the technology. *New Genet Soc.* 2019;38(2):119-141.

San-Millán M, Rissech C, Turbón D. Shape variability of the adult human acetabulum and acetabular fossa related to sex and age by geometric morphometrics. Implications for adult age estimation. *Forensic Sci Int.* 2017, 272: 50–63.

Sarda S, Hannehalli S. Orphan CpG islands as alternative promoters. *Transcription.* 2018;9(3):171-176.

Schmeling A, Garamendi PM, Prieto JL, Landa MI. Forensic Age Estimation in Unaccompanied Minors and Young Living Adults. In: Vieira DN, editor. Forensic medicine -from old problems to new challenges InTech, Rijeka; 2011. pp. 77-120.

Schaefer MC, Black S, Scheuer JL. Juvenile osteology- A laboratory and field manual. London: Elsevier; 2009. Page | 343

Schübeler D. Function and information content of DNA methylation. *Nature*. 2015;517(7534):321-6. doi: 10.1038/nature14192.

Shi L, Jiang F, Ouyang F, Zhang J, Wang Z, Shen X. DNA methylation markers in combination with skeletal and dental ages to improve age estimation in children. *Forensic Sci Int Genet*. 2018;33:1-9. doi: 10.1016/j.fsigen.2017.11.005.

Singh C, Singal K. Teeth as a Tool for Age Estimation: A Mini Review. *J Forensic Sci & Criminal Inves*. 2017;6(3):1-4.

Silva DSBS, Antunes J, Balamurugan K, Duncan G, Alho CS, McCord B. Evaluation of DNA methylation markers and their potential to predict human aging. *Electrophoresis*. 2015;36,1775–1780.

Sliker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin*. 2018;11(25):1-11.

Spólnicka M, Pośpiech E, Pepłońska B, Zbieć-Piekarska R, Makowska Ż, Pięta A, et al. DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups. *Int J Legal Med*. 2017;132(1):1-11.

T

Tan Q, Heijmans BT, Hjelmborg JV, Soerensen M, Christensen K, Christiansen L. Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. *Int J Epidemiol*. 2016;45:1146–1158.

Takasaki T, Tsuji A, Ikeda N, Ohishi M. Age estimation in dental pulp DNA based on human telomere shortening. *Int J Legal Med*. 2003;117:232–234. doi:10.1007/s00414-003-0376-5.

Teixeira F, Cunha E. Aging the elderly: Does the skull tell us something about age at death? In: Algee-Hewitt B, Kim J, editors. *Remodeling Forensic Skeletal Age*. Academic press. In press. 2021.

Thiede C, Prange-Krex G, Freiberg-Richter J, Bornhauser M, Ehninger G. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. *Bone Marrow Transplant*. 2000; 25:575-77.

Thermo Fisher Scientific. User Guide: Quantifiler Human DNA Quantification Kit Quantifiler, Quantifiler Y Human Male DNA Quantification Kit. 2018. Available from <https://www.thermofisher.com/order/catalog/product/4343895#/4343895>

Thong Z, Chan XLS, Tan JYY, Loo ES, Syn CKC. Evaluation of DNA methylation-based age prediction on blood. *Forensic Sci Int Genet Suppl Ser*. 2017;6:e249-e251.

Tsuji A, Ishiko A, Takasaki T, Ikeda N. Estimating age of humans based on telomere shortening. *Forensic Sci Int*. 2002; 126:197–199.

U

Ubelaker DH, Khosrowshahi H. Estimation of age in forensic anthropology: historical perspective and recent methodological advances. *Forensic Sci Res*. 2019;4(1): 1-9. DOI: 10.1080/20961790.2018.1549711.

Ubelaker DH, Shamlou A, Kunkle AE. Forensic anthropology in the global investigation of humanitarian and human rights abuse: Perspective from the published record. *Sci Justice*. 2019a;59(2):203-209. doi: 10.1016/j.scijus.2018.10.008.

Ubelaker DH, Cordero QR, Linton NF. Recent research in Forensic Anthropology. *Eur J Anat* 2020;24 (3): 221-227.

V

Vidaki A, Daniel B, Court DS. Forensic DNA methylation profiling – potential opportunities and challenges. *Forensic Sci Int Genet*. 2013;7:499–507. <http://dx.doi.org/10.1016/j.fsigen.2013.05.004>.

Vidaki A, Giangasparo F, Syndercombe Court D. Discovery of potential DNA methylation markers for forensic tissue identification using bisulphite pyrosequencing. *Electrophoresis*. 2016;37(21):2767-2779.

Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. *Genome Biol*. 2017;18:238-250. Page | 345

Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci Int Genet*. 2017;28:225–236.

Vidaki A, Kayser M. Recent progress, methods and perspectives in forensic epigenetics. *Forensic Sci Int Genet*. 2018. <https://doi.org/10.1016/j.fsigen.2018.08.008>.

W

Waddington CH. The epigenotype. *Endeavour*; 1942(1): 18.

Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jöckel KH, Erbel R, Mühleisen TW, Zenke M, Brümmendorf TH, Wagner W. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol*. 2014;15(2):R24. doi: 10.1186/gb-2014-15-2-r24.

Williams G. The emerging field of forensic epigenetics. *Forensic Sci Int*. 2018;290:e24-e25. doi: 10.1016/j.forsciint.2018.07.019.

Wochna K, Bonikowski R, Śmigielski J, Berent J. Aspartic acid racemization of root dentin used for dental age estimation in a Polish population sample. *Forensic Sci Med Pathol*. 2018;14(3):285-294. doi: 10.1007/s12024-018-9984-8.

X

Xiao FH, Kong QP, Perry B, He YH. Progress on the role of DNA methylation in aging and longevity. *Brief Funct Genomics* 2016;15,454–459. doi: 10.1093/bfgp/elw009.

Xu C, Qu H, Wang G, Xie B, Shi Y, Yang Y, Zhao Z, Hu L, Fang X, Yan J, Feng L. A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Nature*. 2015;1-10.

Y

Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, Zheng SC, Widschwendter M, Beck S, Teschendorff AE. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* 2016;17:205.

Page | 346

Youn A, Wang S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. *Epigenetics.* 2018;13:192-206.

Z

Zampieri M, Ciccarone F, Calabrese R, Franceschi C, Bürkled A, Caiafa P. Reconfiguration of DNA methylation in aging. *Mech Ageing Dev.* 2015;151:60–70.

Zapico SC, Stone-Gordon R, Adserias-Garriga J. The evolution of methodology in biochemical age estimation. In: Adserias-Garriga J, editor. *Age estimation: A multidisciplinary approach.* Elsevier; 2019. pp. 189-197.

Zapico SC, Ubelaker DH. mtDNA Mutations and their role in aging, diseases, and forensic sciences. *Aging Dis.* 2013;4(6):364–380. <https://doi.org/10.14336/AD.2013.0400364>.

Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Ż, Spas A, Parys-Proszek A, Kucharczyk K, Płoski R, Branicki W. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet.* 2015a;14:161–167.

Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Ż, Pałeczka A, Kucharczyk K, Płoski R, Branicki W. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci Int Genet.* 2015b;17:173–179.

Zhang Y, Hapala J, Brenner H, Wagner W. Individual CpG sites that are associated with age and life expectancy become hypomethylated upon aging. *Clin Epigenetics.* 2017a;9:9. doi: 10.1186/s13148-017-0315-9.

Zhang Y, Wilson R, Heiss J, Breitling LP, Saum KU, Schöttker B, Holleczeck B, Waldenberger M, Peters A, Brenner H. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat Commun.* 2017b;8:14617.

Zolotareno AD, Chekalin EV, Bruskin SA. Modern Molecular Genetic Methods for Age Estimation in Forensics. *Russ J Genet.* 2019;55:1460–1471. <https://doi.org/10.1134/S1022795419120147>.

Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM, Driessen GJ, van Dongen JJ, Kayser M, Langerak AW. Estimating human age from T-cell DNA rearrangements. *Current Biol.* 2010;20: R970-R971. Page | 347

Zubakov D, Liu F, Kokmeijer I, Choi Y, van Meurs JBJ, van Ijcken WFJ, Uitterlinden AG, Hofman A, Broer L, van Duijn CM, Lewin J, Kayser M. Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Forensic Sci Int Genet.* 2016;24:33–43.

Zupanič Pajnič I, Debska M, Gornjak Pogorelc B, Vodopivec Mohorčič K, Balažič J, Zupanc T, Štefanič B, Geršak K. Highly efficient automated extraction of DNA from old and contemporary skeletal remains. *J Forensic Leg Med.* 2016;37:78-86. doi: 10.1016/j.jflm.2015.11.001.

Annex I

Forms (in Portuguese)



Requerente/Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com
Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colheita de amostras de sangue, ossos e dentes (Corpos Doados à Ciência)

Faculdade de Medicina da Universidade de:

Coimbra

Porto

Lisboa

-Nº **mínimo** de colheitas 20 indivíduos (F/M)

-Recolha pela investigadora (Helena Correia Dias) na respectiva Faculdade de Medicina, quando existam amostras que justifiquem a deslocação.

Sangue:

-Colheita de 2 mL de sangue periférico ou cardíaco (tubos com EDTA fornecidos)

-Armazenamento: sem centrifugar, congelar a -20°C

Ossos (punch bone ou fragmento do osso)

-tíbia

-clavícula

Fazer 5 a 7 punch bone----- congelar -80°C

Se for fragmento, envolver em papel de alumínio e congelar -80 °C

Dente

- 1º ou 2º Molar (superior ou inferior) saudáveis e sem tratamentos dentários;

-Armazenamento: recipiente fornecido, temperatura ambiente.

ID Corpo Doador à Ciência	Idade	Sexo F/M	Sangue	Osso	Dente	Datas
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___
			Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Sim <input type="checkbox"/> Não <input type="checkbox"/>	Óbito ___/___/___ Colheita da amostra ___/___/___

ID Corpo Doado à Ciência	Observações: Causa da morte e/ou possíveis doenças <i>antemortem</i> conhecidas, ou outras condições clínicas Associadas

Buccal swab consent



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Faculdade de Ciências e Tecnologia
Universidade de Coimbra

UNIVERSIDADE DE COIMBRA

A estimativa da idade é uma questão muito importante nos dias de hoje. É necessária para identificação do indivíduo e para resolução de certas situações civis e judiciais. Com o aumento considerável de imigrantes e refugiados, sem documentos válidos, torna-se importante a criação de novas metodologias para estimar a idade destes indivíduos.

A recolha de saliva para estudos de estimativa da idade é um procedimento indolor e bastante rápido. A investigação será feita no Departamento de Ciências da Vida da Universidade de Coimbra pela investigadora Maria Helena Correia Dias e está incluída no âmbito do seu trabalho de Doutoramento em Antropologia Forense.

A própria investigadora fará a recolha das amostras dos doadores voluntários, após consentimento informado. A investigadora disponibiliza-se a esclarecer qualquer dúvida ou questão, quer no acto da colheita da amostra (esfregaço bucal), quer posteriormente, a qualquer altura (helenacorreiadias30@gmail.com; 962903987).

Para ser incluído neste estudo, apenas necessita de concordar com o respectivo consentimento abaixo exposto.

Consentimento informado

Eu, _____,
autorizo a investigadora Maria Helena Correia Dias, estudante de Doutoramento em Antropologia Forense na Faculdade de Ciências e Tecnologia da Universidade de Coimbra, a recolher um esfregaço bucal para fins de investigação científica incluída no seu trabalho de Doutoramento.

A minha participação é voluntária e tomei conhecimento de que a amostra apenas seria usada no âmbito de investigação científica, sendo que toda a informação obtida a partir da mesma será mantida de forma confidencial e em anonimato.

Tomei conhecimento que os resultados provenientes deste estudo serão publicados cientificamente, salvaguardando-se, contudo, a total confidencialidade das informações dos doadores. Os estudos não terão qualquer tipo de fins lucrativos e a recolha dos esfregaços bucais não provoca qualquer dor ou consequência futura danosa ou prejudicial para o doador.

Page | 356

Declaro que compreendi toda a informação anteriormente exposta neste consentimento, ou explicada pessoalmente pela investigadora, aceitando por isso participar neste estudo de forma voluntária.

Doador: _____

Investigadora: _____

_____, ____ de _____ de ____

Muito obrigada pela participação neste estudo.

Dados do doador:

Nome: _____

Idade: ____

Sexo: ____

Nacionalidade: _____

Annex II
Biological features of Bodies Donated to
Science (BDS)

Biological features of Bodies Donated to Science (BDS)

ID BDS	Age	Sex	Date at death	Type of Sample	Date collection	City of collection
BDS 1	75	M	28/04/2015	Bone	15/05/2018	Coimbra
BDS 2	65	F	24/02/2015	Bone	15/05/2018	Coimbra
BDS 3	65	F	28/07/2016	Bone	15/05/2018	Coimbra
BDS 4	87	F	03/12/2014	Bone	15/05/2018	Coimbra
BDS 5	75	F	28/11/2015	Bone; tooth	19/07/2018	Porto
BDS 6	88	M	01/08/2018	Blood; tooth	03/08/2018	Porto
BDS 7	72	M	05/08/2018	Blood; tooth	07/08/2018	Porto
BDS 8	66	F	10/08/2018	Tooth	31/08/2018	Porto
BDS 9	66	M	24/08/2018	Blood	29/08/2018	Porto
BDS 10	55	F	02/09/2018	Blood	03/09/2018	Porto
BDS 11	84	M	14/09/2018	Blood	17/09/2018	Porto
BDS 12	61	M	22/09/2018	Blood	24/09/2018	Porto
BDS 13	84	M	30/09/2018	Blood	02/10/2018	Porto
BDS 14	49	F	21/10/2018	Blood; tooth Bone	23/10/2018 11/02/2019	Porto
BDS 15	72	F	26/10/2018	Blood; tooth	29/10/2018	Porto
BDS 16	52	F	31/10/2015	Bone	22/11/2018	Porto
BDS 17	77	F	13/11/2015	Bone	22/11/2018	Porto
BDS 18	56	M	17/11/2015	Bone	22/11/2018	Porto
BDS 19	75	F	28/11/2015	Bone	22/11/2018	Porto
BDS 20	79	M	05/04/2016	Bone	22/11/2018	Porto
BDS 21	87	F	12/05/2016	Bone	22/11/2018	Porto
BDS 22	63	M	13/05/2016	Bone	22/11/2018	Porto
BDS 23	54	F	03/06/2016	Bone	22/11/2018	Porto
BDS 24	81	M	01/08/2016	Bone	11/02/2019	Porto
BDS 25	73	F	29/08/2016	Bone	11/02/2019	Porto
BDS 26	79	M	20/09/2016	Bone	11/02/2019	Porto
BDS 27	85	M	30/09/2016	Bone	11/02/2019	Porto
BDS 28	80	M	06/11/2016	Bone	11/02/2019	Porto
BDS 29	93	M	23/02/2017	Bone	11/02/2019	Porto
BDS 30	76	M	27/03/2017	Bone	11/02/2019	Porto
BDS 31	57	M	8/12/2018	Blood	11/12/2018	Porto
BDS 32	81	M	18/12/2018	Blood	20/12/2018	Porto
BDS 33	66	M	30/12/2018	Blood	02/01/2019	Porto
BDS 34	93	F	14/01/2019	Blood	15/01/2019	Porto
BDS 35	82	M	30/01/2019	Tooth	01/02/2019	Porto
BDS 36	48	F	02/02/2019	Blood; tooth	04/02/2019	Porto
BDS 37	80	F	29/03/2016	Bone	16/04/2019	Porto

Annex III
Ethical consents (in Portuguese)

**Ethical Committee of Faculdade de Medicina
da Universidade de Coimbra**

FACULDADE DE MEDICINA
UNIVERSIDADE DE COIMBRA

COMISSÃO DE ÉTICA DA FMUC

Of. Refª **038-CE-2017**

Data 26/06/2017

Page | 363

C/C aos Exmos. Senhores
Investigadores e co-investigadores

Exmo. Senhor
Prof. Doutor Duarte Nuno Vieira
Director da Faculdade de Medicina de
Universidade de Coimbra

Assunto: Pedido de parecer à Comissão de Ética - Projeto de Investigação autónomo (refª CE-034/2017).

Investigador(a) Principal: Maria Helena Correia Dias

Co-Investigador(es): Eugénia Maria Guedes Pinto Antunes da Cunha, Licínio Manuel Mendes Manco e Francisco Manuel Andrade Corte-Real Gonçalves

Título do Projeto: "Metilação de DNA como um preditor de idade à morte em indivíduos mortos e restos humanos".

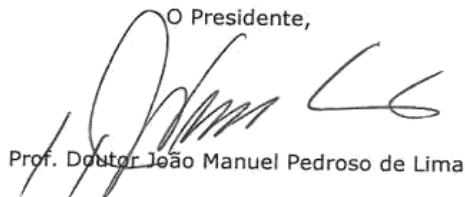
A Comissão de Ética da Faculdade de Medicina, após análise do projeto de investigação supra identificado, decidiu emitir o parecer que a seguir se transcreve:

"Parecer favorável não se excluindo, no entanto, a necessidade de submissão à Comissão de Ética, caso exista, da(s) Instituição(ões) onde será realizado o Projeto".

Queira aceitar os meus melhores cumprimentos.



O Presidente,



Prof. Doutor João Manuel Pedroso de Lima

HC

SERVIÇOS TÉCNICOS DE APOIO À GESTÃO - STAG • COMISSÃO DE ÉTICA

Pólo das Ciências da Saúde • Unidade Central

Azinhaga de Santa Comba, Celas, 3000-354 COIMBRA • PORTUGAL

Tel.: +351 239 857 707 (Ext. 542707) | Fax: +351 239 823 236

E-mail: comissaoetica@fmed.uc.pt | www.fmed.uc.pt

Annex IV

Protocols

Protocol I_DNA extraction from blood samples (full protocol)

Genomic DNA extraction from blood samples performed using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany), according to the instructions of the manufacture:

- 1.** Addition of 20 µl of proteinase K (Qiagen) (10 mg/ml) and 200 µl of Lysis Buffer AL to 200 µl of blood in a 1.5 ml microcentrifuge tube. Mix by vortex.
- 2.** Incubation at 56 °C (10 min). Quick centrifugation.
- 3.** Addition of 200 µl of ethanol (96-100%), vortex (15 sec) and quick centrifugation.
- 4.** Transfer the sample mixture to QIAamp Mini spin column (in a 2 ml collection tube), centrifugation at 8000 rpm (1 min) and remove the collection tube. Replace the collection tube.
- 5.** Pipet 500 µl of Buffer AW1. Centrifugation at 8000 rpm (1 min) and remove the flow-through.
- 6.** Pipet 500 µl of Buffer AW2 and centrifugation at 14.000 rpm (3 min).
- 7.** Put the QIAamp Mini spin column in a new collection tube (2 ml) and centrifuge (1 min) at 14.00 rpm.
- 8.** Use a clean 1.5 ml microcentrifuge tube, put the QIAamp Mini spin column, pipet 200 µl Buffer AE, incubation at room temperature during 5 min for increase the amount of DNA. Centrifugation at 8000 rpm (1 min).
- 9.** Storage the DNA at 4°C or -10°C.

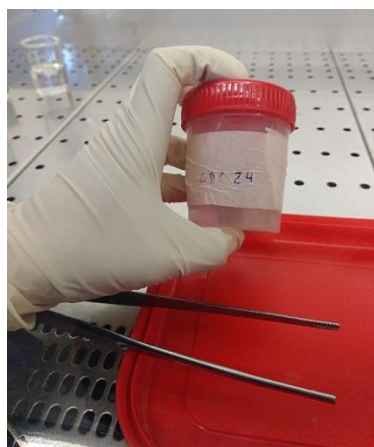
Protocol II_Guide for processing human remains and extract DNA
(step by step with illustrations)

Page | 368

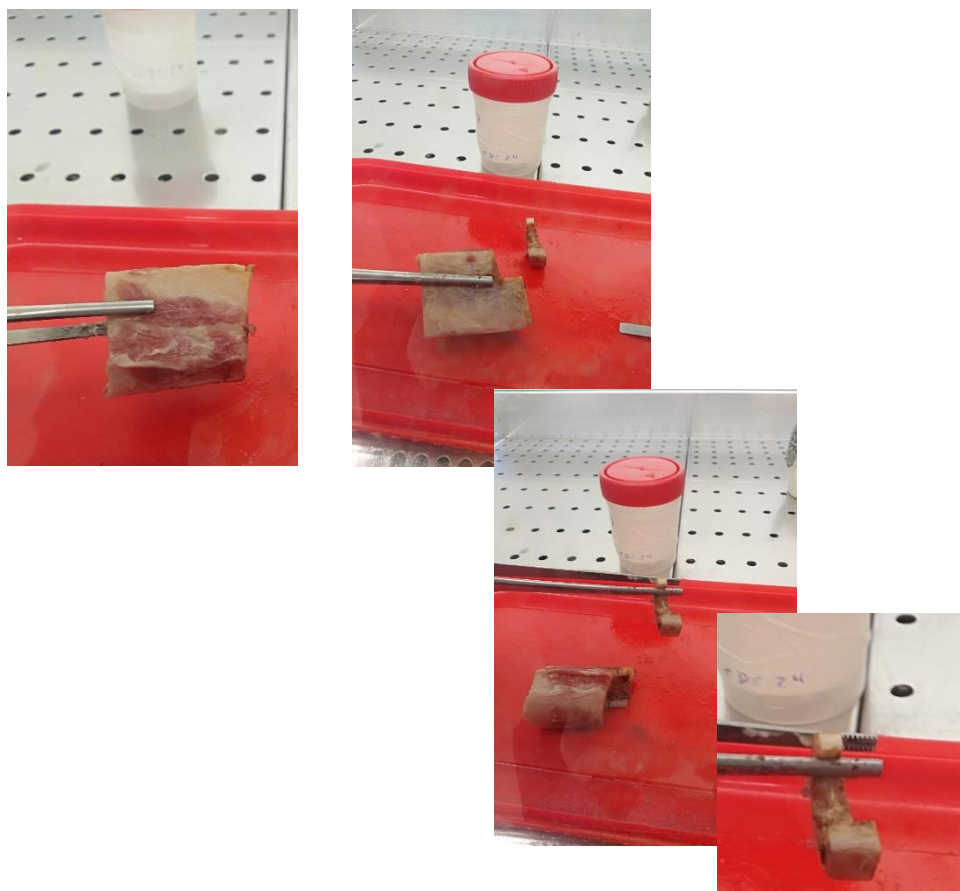
Workstation I at INMLCF: after at least 20 minutes of UV decontamination (5° steps)



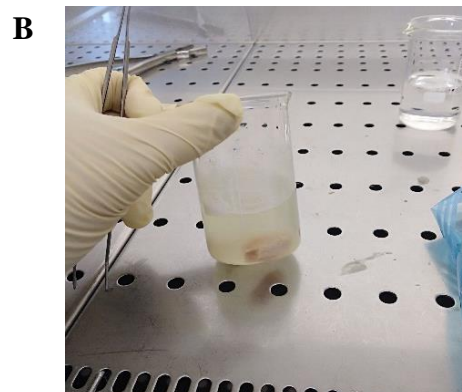
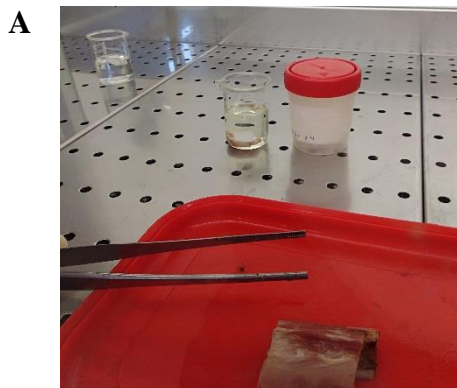
Step 1: After collecting the fresh bone, a new ID label is given.
In example, BDS 24 (CDC 24, *Corpo Doado à Ciência 24*) is a fresh tibial right bone from a male with 81 years old, collected in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*.

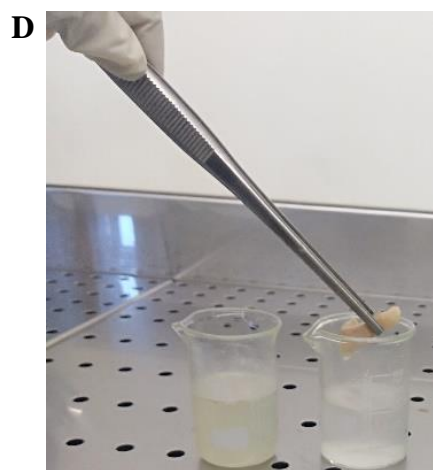
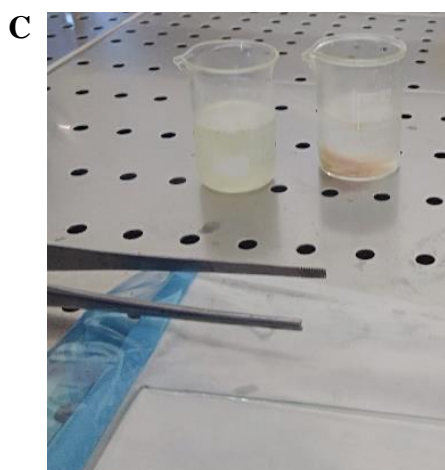


Step 2: Bone fragment cut (before cleaning the bone).



Step 3: Clean the bone fragment chemically with commercial bleach during 5 minutes (A, B), then put in distilled water (ddH₂O) during 5 minutes (C, D).



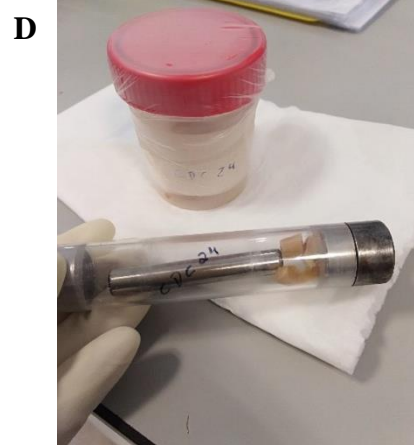
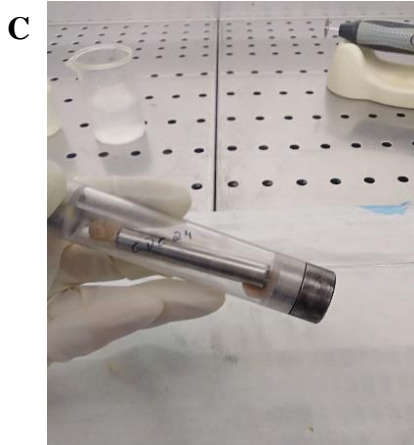


Step 4: Clean the bone mechanically with a drill (physical removal of the surface using a rotary sanding tool).



Step 5: Make some cuts (around 0.5 x 0.5 cm) in the bone fragment (A and B) and put in a vial for grinding bone (C and D).

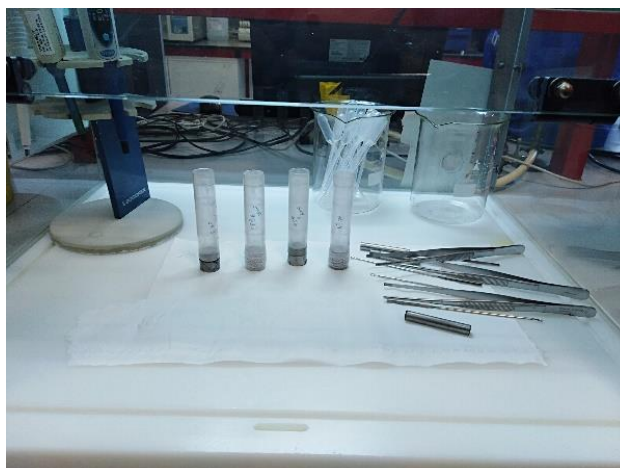




Grinding bones in SPEX Sample Prep Freezer/Mill 6770 with liquid nitrogen.



In a workstation II in INMLCF: after at least 20 minutes of UV decontamination (3° steps)



Step 1: around 50 mg of fine powder were put in a lysis tube (standard protocol). Prepared the PrepFiler BTA™ lysis solution in workstation II. Put the PrepFiler Bone Lysate Tube in a thermal shaker and incubated at 56°C, 750 rpm overnight.



Step 2: DNA extraction using a semi-automatic protocol with *PrepFiler Express BTA™ Forensic DNA Extraction kit* (Thermo Fisher Scientific Inc., Waltham, MA, USA).



Protocol III_DNA extraction from buccal swabs (full protocol)

Genomic DNA from buccal swabs was extracted using the FavorPrep™ Genomic DNA mini kit (Favorgen Biotech Corp, Taiwan), according to the instructions of the manufacture:

- 1.** Resuspending the buccal swab in 1 ml of dH₂O, centrifuged at 12.000 rpm (3 min) and discard the supernatant.
- 2.** Addition of 200 µl of FABG Buffer and 20 µl of proteinase K (10 mg/ml). Vortex.
- 3.** Put in the Thermo mix at 56°C (at least one hour).
- 4.** Preheat the elution buffer in a 70°C water bath.
- 5.** Addition of 200 µl of ethanol (100%) and vortex (10 sec).
- 6.** Transfer the sample mixture to FABG column (with a 2 ml collection tube), centrifuge at 12.000 rpm (5 min) and remove the flow-through.
- 7.** Wash FABG column with 400 µl of W1 buffer, centrifuge at 12.000 rpm (1 min) and remove the flow-through.
- 8.** Add 600 µl of Wash Buffer (ethanol added) to the FABG column. Centrifugation at 12.000 rpm (1 min), remove the flow-through.
- 9.** Dry the column with a final centrifugation during 3 min at 12.000 rpm.
- 10.** For DNA elution, put the FABG column to a new 1.5 ml microcentrifuge tube, add 60 µl of preheated elution buffer to the column and centrifuge at 12.000 rpm (1 min). Isolated DNA can be stored at 4°C or -10°C.

Annex V
Supplementary Material Tables

A. DNA methylation age estimation in blood samples

Table S1: Univariate linear regression analysis of the 37 CpG sites in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* loci in 53 blood samples from living individuals.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 53)	CpG1	Chr6: 11044628	0.920	0.846	0.843	11.61	2.14 × 10⁻²²
	CpG2	Chr6: 11044631	0.868	0.753	0.749	14.71	3.95 × 10⁻¹⁷
	CpG3	Chr6: 11044634	0.884	0.781	0.777	13.85	1.82 × 10⁻¹⁸
	CpG4	Chr6: 11044640	0.916	0.838	0.835	11.91	8.22 × 10⁻²²
	CpG5	Chr6: 11044642	0.936	0.876	0.873	10.43	9.22 × 10⁻²⁵
	<u>CpG6</u>	Chr6: 11044644	0.936	0.877	0.874	10.41	7.97 × 10⁻²⁵
	CpG7	Chr6: 11044647	0.914	0.836	0.833	11.99	1.15 × 10⁻²¹
	CpG8	Chr6: 11044655	0.850	0.722	0.716	15.63	8.89 × 10⁻¹⁶
	CpG9	Chr6: 11044661	0.860	0.740	0.735	15.10	1.53 × 10⁻¹⁶
<i>FHL2</i> (N = 53)	CpG1	Chr2:105399282	0.937	0.878	0.876	10.33	5.68 × 10⁻²⁵
	CpG2	Chr2: 105399288	0.940	0.883	0.881	10.12	1.89 × 10⁻²⁵
	<u>CpG3</u>	Chr2: 105399291	0.940	0.884	0.881	10.11	1.78 × 10⁻²⁵
	CpG4	Chr2: 105399297	0.814	0.662	0.655	17.22	1.31 × 10⁻¹³
	CpG5	Chr2: 105399300	0.921	0.849	0.846	11.53	1.48 × 10⁻²²
	CpG6	Chr2: 105399310	0.848	0.719	0.714	15.70	1.11 × 10⁻¹⁵
	CpG7	Chr2: 105399314	0.263	0.069	0.051	28.59	0.057
	CpG8	Chr2: 105399316	0.530	0.281	0.267	25.12	0.00004
	CpG9	Chr2: 105399323	-0.079	0.006	-0.013	29.54	0.575
	CpG10	Chr2: 105399327	-0.262	0.068	0.050	28.60	0.058
	CpG11	Chr2: 105399338	-0.794	0.631	0.623	18.01	1.29 × 10⁻¹²
	CpG12	Chr2: 105399340	-0.658	0.432	0.421	22.32	8.82 × 10⁻⁸
<i>EDARADD</i> (N = 53)	CpG1	Chr1: 236394458	0.014	0.000	-0.019	29.63	0.920
	CpG2	Chr1: 236394441	-0.745	0.555	0.546	19.76	1.59 × 10⁻¹⁰
	<u>CpG3</u>	Chr1: 236394382	-0.888	0.788	0.784	13.63	7.86 × 10⁻¹⁹
	CpG4	Chr1: 236394370	-0.619	0.383	0.371	23.27	7.73 × 10⁻⁷
<i>PDE4C</i> (N = 53)	CpG1	Chr19: 18233139	0.849	0.720	0.715	15.67	1.02 × 10⁻¹⁵
	<u>CpG2</u>	Chr19: 18233133	0.852	0.725	0.720	15.53	6.32 × 10⁻¹⁶
	CpG3	Chr19: 18233131	0.827	0.684	0.678	16.64	2.25 × 10⁻¹⁴
	CpG4	Chr19: 18233127	0.525	0.276	0.261	25.22	0.000055
	CpG5	Chr19: 18233105	0.804	0.647	0.640	17.61	4.07 × 10⁻¹³
	CpG6	Chr19: 18233091	-0.432	0.187	0.171	26.72	0.001
	CpG7	Chr19: 18233082	0.140	0.020	0.000	29.34	0.318
	CpG8	Chr19: 18233079	-0.273	0.074	0.056	28.51	0.048
	CpG9	Chr19: 18233070	-0.410	0.168	0.152	27.02	0.002
	CpG10	Chr19: 18233058	-0.097	0.009	-0.010	29.49	0.489
	CpG11	Chr19: 18233048	-0.116	0.013	-0.006	29.43	0.409
	CpG12	Chr19: 18233042	0.015	0.000	-0.019	29.62	0.916

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S2: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* in blood samples from living individuals.

Marker	Coefficient	P-value
(Intercept)	39.280	0.000
<i>ELOVL2</i>	63.817	0.001
<i>FHL2</i>	70.186	0.001
<i>C1orf132</i>	-47.768	0.000
<i>TRIM59</i>	-1.100	0.948
<i>KLF14</i>	30.324	0.479

Abbreviations: Significant p-values are in bold.

Table S3: Univariate linear regression analysis of the 43 CpG sites in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* loci in 51 blood samples from deceased individuals.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 49)	CpG1	Chr6: 11044628	0.781	0.610	0.602	10.28	3.50 × 10⁻¹¹
	CpG2	Chr6: 11044631	0.700	0.490	0.479	11.76	2.15 × 10⁻⁸
	CpG3	Chr6: 11044634	0.740	0.547	0.538	11.09	1.25 × 10⁻⁹
	<u>CpG4</u>	Chr6: 11044640	0.785	0.617	0.608	10.20	2.39 × 10⁻¹¹
	CpG5	Chr6: 11044642	0.700	0.490	0.479	11.76	2.16 × 10⁻⁸
	CpG6	Chr6: 11044644	0.764	0.584	0.575	10.63	1.65 × 10⁻¹⁰
	CpG7	Chr6: 11044647	0.758	0.575	0.566	10.74	2.82 × 10⁻¹⁰
	CpG8	Chr6: 11044655	0.706	0.498	0.488	11.67	1.48 × 10⁻⁸
	CpG9	Chr6: 11044661	0.663	0.440	0.428	12.33	2.08 × 10⁻⁷
<i>FHL2</i> (N = 51)	CpG1	Chr2:105399282	0.431	0.186	0.169	14.64	0.002
	<u>CpG2</u>	Chr2: 105399288	0.465	0.216	0.200	14.36	0.000584
	CpG3	Chr2: 105399291	0.459	0.211	0.195	14.41	0.000704
	CpG4	Chr2: 105399297	0.141	0.020	0.000	16.06	0.323
	CpG5	Chr2: 105399300	0.388	0.151	0.133	14.95	0.005
	CpG6	Chr2: 105399310	0.224	0.050	0.031	15.81	0.114
	CpG7	Chr2: 105399314	0.118	0.014	-0.006	16.11	0.411
	CpG8	Chr2: 105399316	0.194	0.038	0.018	15.92	0.173
	CpG9	Chr2: 105399323	-0.031	0.001	-0.019	16.22	0.830
	CpG10	Chr2: 105399327	-0.037	0.001	-0.019	16.21	0.796
	CpG11	Chr2: 105399338	-0.214	0.046	0.026	15.85	0.131
	CpG12	Chr2: 105399340	-0.158	0.025	0.005	16.02	0.269
<i>EDARADD</i> (N = 51)	CpG1	Chr1: 236394458	0.001	0.000	-0.020	16.22	0.993
	CpG2	Chr1: 236394441	-0.351	0.123	0.105	15.19	0.012
	<u>CpG3</u>	Chr1: 236394382	-0.621	0.385	0.373	12.72	0.000001
	CpG4	Chr1: 236394370	-0.321	0.103	0.085	15.37	0.022
<i>PDE4C</i> (N = 49)	CpG1	Chr19: 18233139	0.401	0.161	0.143	15.10	0.004
	<u>CpG2</u>	Chr19: 18233133	0.592	0.350	0.336	13.28	0.000008
	CpG3	Chr19: 18233131	0.585	0.342	0.328	13.36	0.000010
	CpG4	Chr19: 18233127	0.231	0.053	0.033	16.03	0.111
	CpG5	Chr19: 18233105	0.299	0.089	0.070	15.72	0.037
	CpG6	Chr19: 18233091	-0.335	0.112	0.093	15.52	0.019
	CpG7	Chr19: 18233082	-0.169	0.029	0.008	16.24	0.246
	CpG8	Chr19: 18233079	-0.199	0.040	0.019	16.15	0.170
	CpG9	Chr19: 18233070	-0.340	0.116	0.097	15.49	0.017
	CpG10	Chr19: 18233058	-0.248	0.061	0.041	15.96	0.086
	CpG11	Chr19: 18233048	-0.159	0.025	0.005	16.27	0.275
	CpG12	Chr19: 18233042	-0.267	0.071	0.051	15.88	0.064
<i>C1orf132</i> (N = 49)	<u>CpG1</u>	Chr1: 207823681	-0.634	0.402	0.389	12.68	0.000001
	CpG2	Chr1: 207823675	-0.491	0.241	0.225	14.27	0.000338
	CpG3	Chr1: 207823672	-0.570	0.325	0.311	13.47	0.000019
	CpG4	Chr1: 207823660	-0.451	0.203	0.186	14.63	0.001
	CpG5	Chr1: 207823657	-0.427	0.183	0.165	14.82	0.002
	CpG6	Chr1: 207823637	-0.498	0.248	0.232	14.21	0.000268

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S4: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* in blood samples from deceased individuals.

Marker	Coefficient	P-value
(Intercept)	18.924	0.029
<i>ELOVL2</i>	65.369	0.005
<i>FHL2</i>	36.006	0.032
<i>C1orf132</i>	-26.255	0.001
<i>TRIM59</i>	43.886	0.014
<i>KLF14</i>	40.717	0.424

Abbreviations: Significant p-values are in bold.

Table S5: Univariate linear regression analysis of the 37 CpG sites in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* loci in 144 blood samples from living and deceased individuals.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 141)	CpG1	Chr6: 11044628	0.872	0.760	0.759	12.28	5.88 × 10⁻⁴⁵
	CpG2	Chr6: 11044631	0.794	0.631	0.628	15.25	7.33 × 10⁻³²
	CpG3	Chr6: 11044634	0.859	0.738	0.736	12.84	2.79 × 10⁻⁴²
	CpG4	Chr6: 11044640	0.880	0.774	0.772	11.93	9.75 × 10⁻⁴⁷
	CpG5	Chr6: 11044642	0.839	0.704	0.702	13.65	1.46 × 10⁻³⁸
	<u>CpG6</u>	Chr6: 11044644	0.892	0.796	0.795	11.33	7.77 × 10⁻⁵⁰
	CpG7	Chr6: 11044647	0.876	0.767	0.765	12.10	7.98 × 10⁻⁴⁶
	CpG8	Chr6: 11044655	0.758	0.574	0.571	16.38	1.54 × 10⁻²⁷
	CpG9	Chr6: 11044661	0.745	0.554	0.551	16.75	3.66 × 10⁻²⁶
<i>FHL2</i> (N = 144)	<u>CpG1</u>	Chr2:105399282	0.828	0.686	0.684	13.91	1.41 × 10⁻³⁷
	CpG2	Chr2:105399288	0.821	0.675	0.672	14.17	1.92 × 10⁻³⁶
	CpG3	Chr2:105399291	0.813	0.661	0.659	14.45	3.26 × 10⁻³⁵
	CpG4	Chr2:105399297	0.623	0.399	0.395	19.25	2.05 × 10⁻¹⁷
	CpG5	Chr2:105399300	0.774	0.600	0.597	15.72	5.13 × 10⁻³⁰
	CpG6	Chr2:105399310	0.667	0.445	0.441	18.50	6.84 × 10⁻²⁰
	CpG7	Chr2:105399314	0.346	0.120	0.114	23.30	0.000021
	CpG8	Chr2:105399316	0.460	0.212	0.206	22.05	6.54 × 10⁻⁹
	CpG9	Chr2:105399323	0.187	0.035	0.028	24.40	0.025069
	CpG10	Chr2:105399327	0.056	0.003	-0.004	24.80	0.508107
	CpG11	Chr2:105399338	-0.482	0.232	0.227	21.76	9.59 × 10⁻¹⁰
	CpG12	Chr2:105399340	-0.118	0.014	0.007	24.67	0.160230
<i>EDARADD</i> (N = 143)	CpG1	Chr1:236394458	-0.026	0.001	-0.006	24.90	0.761497
	CpG2	Chr1:236394441	-0.629	0.395	0.391	19.38	4.16 × 10⁻¹⁷
	<u>CpG3</u>	Chr1:236394382	-0.786	0.617	0.615	15.41	3.27 × 10⁻³¹
	CpG4	Chr1:236394370	-0.530	0.281	0.275	21.26	1.03 × 10⁻¹¹
<i>PDE4C</i> (N = 141)	CpG1	Chr19:18233139	0.785	0.617	0.614	15.53	9.25 × 10⁻³¹
	<u>CpG2</u>	Chr19:18233133	0.830	0.689	0.687	13.99	4.66 × 10⁻³⁷
	CpG3	Chr19:18233131	0.813	0.660	0.658	14.63	2.16 × 10⁻³⁴
	CpG4	Chr19:18233127	0.526	0.277	0.272	21.33	2.03 × 10⁻¹¹
	CpG5	Chr19:18233105	0.726	0.527	0.524	17.26	2.33 × 10⁻²⁴
	CpG6	Chr19:18233091	-0.370	0.137	0.131	23.30	0.000006
	CpG7	Chr19:18233082	0.065	0.004	-0.003	25.03	0.445853
	CpG8	Chr19:18233079	-0.234	0.055	0.048	24.39	0.005193
	CpG9	Chr19:18233070	-0.417	0.174	0.168	22.81	2.78 × 10⁻⁷
	CpG10	Chr19:18233058	-0.055	0.003	-0.004	25.05	0.520182
	CpG11	Chr19:18233048	-0.163	0.027	0.020	24.75	0.051747
	CpG12	Chr19:18233042	-0.005	0.000	-0.007	25.09	0.957291

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S6: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* in blood samples from living and deceased individuals.

Page | 384

Marker	Coefficient	P-value
(Intercept)	21.259	0.000
<i>ELOVL2</i>	69.782	0.000
<i>FHL2</i>	45.197	0.000
<i>C1orf132</i>	-32.729	0.000
<i>TRIM59</i>	33.794	0.002
<i>KLF14</i>	41.210	0.186

Abbreviations: Significant p-values are in bold.

Table S7: Statistical parameters obtained in a multiple regression model with the four CpGs in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes in blood samples from deceased individuals.

Marker	Coefficient	P-value
(Intercept)	-40.063	0.120
<i>ELOVL2</i> CpG4	100.423	0.000
<i>FHL2</i> CpG2	59.531	0.007
<i>EDARADD</i> CpG3	-74.494	0.000
<i>PDE4C</i> CpG2	26.479	0.229

Abbreviations: Significant p-values are in bold.

B. DNA methylation age estimation in tooth samples

Table S8: Univariate linear regression analysis of the 43 CpG sites in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* loci in 31 tooth samples from living and deceased individuals.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 31)	CpG1	Chr6: 11044628	0.237	0.056	0.024	19.23	0.199224
	CpG2	Chr6: 11044631	0.280	0.079	0.047	19.00	0.126816
	<u>CpG3</u>	Chr6: 11044634	0.379	0.143	0.114	18.32	0.035738
	CpG4	Chr6: 11044640	0.221	0.049	0.016	19.31	0.233055
	CpG5	Chr6: 11044642	0.375	0.141	0.111	18.35	0.037618
	CpG6	Chr6: 11044644	0.336	0.113	0.082	18.65	0.064411
	CpG7	Chr6: 11044647	0.195	0.038	0.005	19.42	0.292839
	CpG8	Chr6: 11044655	0.126	0.016	-0.018	19.64	0.497705
	CpG9	Chr6: 11044661	0.270	0.073	0.041	19.06	0.141588
<i>FHL2</i> (N = 30)	CpG1	Chr2:105399282	0.451	0.204	0.175	17.26	0.012313
	CpG2	Chr2: 105399288	0.409	0.167	0.137	17.66	0.024895
	CpG3	Chr2: 105399291	0.517	0.267	0.241	16.56	0.003430
	<u>CpG4</u>	Chr2: 105399297	0.658	0.433	0.413	14.57	0.000078
	CpG5	Chr2: 105399300	0.508	0.258	0.231	16.67	0.004167
	CpG6	Chr2: 105399310	0.377	0.142	0.112	17.92	0.039980
	CpG7	Chr2: 105399314	0.311	0.097	0.065	18.39	0.094135
	CpG8	Chr2: 105399316	0.238	0.057	0.023	18.79	0.205182
	CpG9	Chr2: 105399323	0.150	0.022	-0.012	19.13	0.429448
	CpG10	Chr2: 105399327	0.063	0.004	-0.032	19.31	0.742681
	CpG11	Chr2: 105399338	-0.330	0.109	0.077	18.26	0.074968
	CpG12	Chr2: 105399340	-0.092	0.008	-0.027	19.26	0.629751
<i>EDARADD</i> (N = 30)	CpG1	Chr1: 236394458	0.012	0.000	-0.036	19.32	0.950015
	CpG2	Chr1: 236394441	0.137	0.019	-0.016	19.14	0.470412
	CpG3	Chr1: 236394382	-0.124	0.015	-0.020	19.17	0.513074
	CpG4	Chr1: 236394370	-0.182	0.033	-0.001	19.00	0.336123
<i>PDE4C</i> (N = 27)	<u>CpG1</u>	Chr19: 18233139	0.474	0.224	0.193	17.54	0.012588
	CpG2	Chr19: 18233133	0.466	0.217	0.186	17.62	0.014265
	CpG3	Chr19: 18233131	0.391	0.153	0.119	18.33	0.043908
	CpG4	Chr19: 18233127	0.437	0.191	0.158	17.91	0.022783
	CpG5	Chr19: 18233105	0.329	0.108	0.073	18.80	0.093455
	CpG6	Chr19: 18233091	0.258	0.067	0.029	19.24	0.194010
	CpG7	Chr19: 18233082	0.364	0.132	0.098	18.55	0.062155
	CpG8	Chr19: 18233079	0.246	0.061	0.023	19.30	0.215924
	CpG9	Chr19: 18233070	0.062	0.004	-0.036	19.87	0.758269
	CpG10	Chr19: 18233058	0.167	0.028	-0.014	19.06	0.424049
	CpG11	Chr19: 18233048	0.238	0.057	0.009	19.65	0.286645
	CpG12	Chr19: 18233042	0.290	0.084	0.038	19.36	0.190559
<i>C1orf132</i> (N = 29)	CpG1	Chr1: 207823681	-0.123	0.015	-0.021	18.68	0.525111
	CpG2	Chr1: 207823675	-0.188	0.035	0.000	18.48	0.328168
	CpG3	Chr1: 207823672	0.104	0.011	-0.026	18.72	0.593103
	CpG4	Chr1: 207823660	0.029	0.001	-0.038	18.89	0.882728
	CpG5	Chr1: 207823657	-0.255	0.065	0.026	18.35	0.208917
	CpG6	Chr1: 207823637	0.126	0.016	-0.022	18.74	0.521969

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S9: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *FHL2*, *ELOVL2* and *PDE4C* in tooth samples from living and deceased individuals.

Page | 388

Marker	Coefficient	P-value
(Intercept)	-115.647	0.039
<i>FHL2</i> CpG4	231.807	0.002
<i>PDE4C</i> CpG1	7.438	0.729
<i>ELOVL2</i> CpG3	3.649	0.947

Abbreviations: Significant p-values are in bold.

Table S10: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *KLF14* and *TRIM59* in tooth samples from living and deceased individuals.

Marker	Coefficient	P-value
(Intercept)	12.126	0.076
<i>ELOVL2</i>	113.705	0.008
<i>TRIM59</i>	-13.768	0.814
<i>KLF14</i>	296.217	0.000

Abbreviations: Significant p-values are in bold.

C. DNA methylation age estimation in bone samples

Table S11: Univariate linear regression analysis of the 43 CpG sites in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* loci in 29 bone samples collected during autopsies.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 29)	CpG1	Chr6: 11044628	0.736	0.542	0.525	9.35	0.000005
	CpG2	Chr6: 11044631	0.628	0.395	0.372	10.75	0.000263
	CpG3	Chr6: 11044634	0.764	0.583	0.568	8.93	0.000001
	CpG4	Chr6: 11044640	0.780	0.609	0.594	8.65	6.02 × 10⁻⁷
	CpG5	Chr6: 11044642	0.777	0.603	0.589	8.71	7.29 × 10⁻⁷
	CpG6	Chr6: 11044644	0.852	0.725	0.715	7.24	4.64 × 10⁻⁹
	CpG7	Chr6: 11044647	0.718	0.515	0.497	9.62	0.000012
	CpG8	Chr6: 11044655	0.717	0.514	0.496	9.63	0.000012
	CpG9	Chr6: 11044661	0.590	0.348	0.324	11.16	0.000750
<i>FHL2</i> (N = 29)	CpG1	Chr2:105399282	0.692	0.479	0.460	9.97	0.000032
	CpG2	Chr2:105399288	0.654	0.428	0.407	10.45	0.000118
	CpG3	Chr2:105399291	0.480	0.230	0.202	12.12	0.008397
	CpG4	Chr2:105399297	0.371	0.138	0.106	12.83	0.047401
	CpG5	Chr2:105399300	0.635	0.403	0.381	10.66	0.000215
	CpG6	Chr2:105399310	0.389	0.151	0.120	12.73	0.037139
	CpG7	Chr2:105399314	0.288	0.083	0.049	13.24	0.130072
	CpG8	Chr2:105399316	0.379	0.144	0.112	12.79	0.042569
	CpG9	Chr2:105399323	0.176	0.031	-0.005	13.60	0.361021
	CpG10	Chr2:105399327	0.009	0.000	-0.037	13.82	0.962069
	CpG11	Chr2:105399338	-0.313	0.098	0.064	13.13	0.098775
	CpG12	Chr2:105399340	-0.004	0.000	-0.037	13.82	0.983379
<i>EDARADD</i> (N = 29)	CpG1	Chr1:236394458	-0.212	0.045	0.010	13.51	0.269687
	CpG2	Chr1:236394441	-0.414	0.171	0.140	12.58	0.025729
	CpG3	Chr1:236394382	-0.561	0.314	0.289	11.45	0.001564
	CpG4	Chr1:236394370	-0.441	0.195	0.165	12.40	0.016516
<i>PDE4C</i> (N = 28)	CpG1	Chr19:18233139	0.592	0.351	0.326	11.34	0.000902
	CpG2	Chr19:18233133	0.690	0.476	0.456	10.19	0.000049
	CpG3	Chr19:18233131	0.620	0.385	0.361	11.04	0.000430
	CpG4	Chr19:18233127	0.490	0.240	0.210	12.27	0.008183
	CpG5	Chr19:18233105	0.633	0.401	0.378	10.89	0.000299
	CpG6	Chr19:18233091	0.238	0.057	0.020	13.66	0.223136
	CpG7	Chr19:18233082	0.287	0.083	0.047	13.47	0.138316
	CpG8	Chr19:18233079	-0.028	0.001	-0.038	14.06	0.887979
	CpG9	Chr19:18233070	-0.174	0.030	-0.007	13.85	0.375355
	CpG10	Chr19:18233058	0.196	0.039	0.002	13.80	0.316687
	CpG11	Chr19:18233048	-0.144	0.021	-0.017	13.92	0.464184
	CpG12	Chr19:18233042	-0.085	0.007	-0.031	14.02	0.666801
<i>C1orf132</i> (N = 29)	CpG1	Chr1:207823681	-0.834	0.695	0.684	7.63	1.93 × 10⁻⁸
	CpG2	Chr1:207823675	-0.753	0.567	0.551	9.09	0.000002
	CpG3	Chr1:207823672	-0.774	0.599	0.585	8.75	8.30 × 10⁻⁷
	CpG4	Chr1:207823660	-0.634	0.402	0.380	10.69	0.000223
	CpG5	Chr1:207823657	-0.311	0.097	0.063	13.14	0.100697
	CpG6	Chr1:207823637	-0.408	0.166	0.136	12.62	0.028010

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S12: Statistical parameters obtained in a multiple regression model with the five CpGs in genes *ELOVL2*, *KLF14*, *TRIM59*, *C1orf132* and *FHL2* in bone samples collected during autopsies.

Marker	Coefficient	P-value
(Intercept)	42.089	0.013
<i>ELOVL2</i>	-7.354	0.899
<i>TRIM59</i>	52.798	0.097
<i>KLF14</i>	120.165	0.040
<i>C1orf132</i>	-35.458	0.038
<i>FHL2</i>	61.184	0.283

Page | 392

Abbreviations: Significant p-values are in bold.

Table S13: Univariate linear regression analysis of the 43 CpG sites in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* loci in 22 bone samples from BDS.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 22)	CpG1	Chr6: 11044628	-0.223	0.050	0.002	12.14	0.317751
	CpG2	Chr6: 11044631	0.154	0.024	-0.025	12.31	0.494247
	CpG3	Chr6: 11044634	-0.006	0.000	-0.050	12.45	0.978553
	CpG4	Chr6: 11044640	-0.077	0.006	-0.044	12.42	0.734888
	CpG5	Chr6: 11044642	-0.003	0.000	-0.050	12.45	0.990470
	CpG6	Chr6: 11044644	-0.059	0.004	-0.046	12.43	0.793437
	CpG7	Chr6: 11044647	-0.056	0.003	-0.047	12.43	0.804966
	<u>CpG8</u>	Chr6: 11044655	0.184	0.034	-0.015	12.24	0.413208
	CpG9	Chr6: 11044661	-0.088	0.008	-0.042	12.41	0.695851
<i>FHL2</i> (N = 22)	CpG1	Chr2:105399282	0.580	0.337	0.303	10.14	0.004648
	CpG2	Chr2: 105399288	0.641	0.411	0.381	9.56	0.001315
	<u>CpG3</u>	Chr2: 105399291	0.655	0.429	0.400	9.41	0.000945
	CpG4	Chr2: 105399297	0.370	0.137	0.093	11.57	0.090455
	CpG5	Chr2: 105399300	0.652	0.425	0.396	9.44	0.001009
	CpG6	Chr2: 105399310	0.584	0.342	0.309	10.11	0.004287
	CpG7	Chr2: 105399314	0.341	0.116	0.072	11.71	0.120096
	CpG8	Chr2: 105399316	0.459	0.210	0.171	11.07	0.031820
	CpG9	Chr2: 105399323	0.198	0.039	-0.009	12.21	0.377918
	CpG10	Chr2: 105399327	0.367	0.135	0.092	11.58	0.092745
	CpG11	Chr2: 105399338	0.062	0.004	-0.046	12.43	0.783816
	CpG12	Chr2: 105399340	0.126	0.016	-0.033	12.36	0.577520
<i>EDARADD</i> (N = 22)	CpG1	Chr1: 236394458	-0.404	0.163	0.121	11.39	0.062444
	CpG2	Chr1: 236394441	-0.282	0.080	0.034	11.95	0.202829
	<u>CpG3</u>	Chr1: 236394382	-0.430	0.185	0.144	11.25	0.045937
	CpG4	Chr1: 236394370	-0.385	0.149	0.106	11.49	0.076444
<i>PDE4C</i> (N = 21)	CpG1	Chr19: 18233139	0.259	0.067	0.018	11.57	0.257413
	CpG2	Chr19: 18233133	0.372	0.138	0.093	11.12	0.097173
	<u>CpG3</u>	Chr19: 18233131	0.771	0.594	0.572	7.63	0.000044
	CpG4	Chr19: 18233127	0.579	0.336	0.301	9.76	0.005922
	CpG5	Chr19: 18233105	0.556	0.309	0.273	9.95	0.008828
	CpG6	Chr19: 18233091	0.115	0.013	-0.039	11.89	0.618456
	CpG7	Chr19: 18233082	0.115	0.013	-0.039	11.89	0.621088
	CpG8	Chr19: 18233079	0.151	0.023	-0.029	11.84	0.512878
	CpG9	Chr19: 18233070	0.240	0.057	0.008	11.63	0.295684
	CpG10	Chr19: 18233058	-0.091	0.008	-0.044	11.92	0.695193
	CpG11	Chr19: 18233048	0.265	0.070	0.021	11.54	0.245144
	CpG12	Chr19: 18233042	-0.151	0.023	-0.032	12.14	0.525183
<i>C1orf132</i> (N = 22)	<u>CpG1</u>	Chr1: 207823681	0.305	0.093	0.048	11.86	0.166966
	CpG2	Chr1: 207823675	0.138	0.019	-0.030	12.34	0.541185
	CpG3	Chr1: 207823672	0.055	0.003	-0.047	12.44	0.806656
	CpG4	Chr1: 207823660	0.247	0.061	0.014	12.07	0.266854
	CpG5	Chr1: 207823657	0.239	0.057	0.010	12.09	0.283866
	CpG6	Chr1: 207823637	0.058	0.003	-0.046	12.43	0.798367

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S14: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *FHL2*, *EDARADD* and *PDE4C* in bone samples from BDS.

Marker	Coefficient	P-value
(Intercept)	19.753	0.652
<i>EDARADD</i> CpG3	-53.625	0.283
<i>PDE4C</i> CpG3	118.822	0.002
<i>FHL2</i> CpG3	27.766	0.505

Abbreviations: Significant p-values are in bold.

D. DNA methylation age estimation in buccal swabs

Table S15: Univariate linear regression analysis of the nine CpG sites in *ELOVL2* locus in 23 buccal swabs.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 23)	<u>CpG1</u>	Chr6: 11044628	0.823	0.677	0.662	14.40	0.000001
	CpG2	Chr6: 11044631	0.533	0.284	0.250	21.45	0.008807
	CpG3	Chr6: 11044634	0.631	0.398	0.370	19.66	0.001241
	CpG4	Chr6: 11044640	0.787	0.619	0.601	15.65	0.000008
	CpG5	Chr6: 11044642	0.806	0.650	0.633	14.99	0.000003
	CpG6	Chr6: 11044644	0.683	0.467	0.441	18.51	0.000328
	CpG7	Chr6: 11044647	0.622	0.387	0.357	19.85	0.001538
	CpG8	Chr6: 11044655	0.649	0.421	0.394	19.29	0.000808
	CpG9	Chr6: 11044661	0.485	0.235	0.199	22.17	0.019061

Page | 397

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S16: Statistical parameters obtained in a multiple regression model with the three CpGs in gene *ELOVL2* in buccal swabs from living individuals.

<i>ELOVL2</i>	Coefficient	P-value
(Intercept)	-83.022	0.000
CpG1	81.881	0.023
CpG4	76.023	0.011
CpG5	11.870	0.656

Abbreviations: Significant p-values are in bold.

Table S17: Statistical parameters obtained in a multiple regression model with the three CpGs in genes *ELOVL2*, *TRIM59* and *KLF14* in buccal swabs from living individuals.

Marker	Coefficient	P-value
(Intercept)	-9.019	0.003
<i>ELOVL2</i>	19.386	0.291
<i>TRIM59</i>	81.446	0.000
<i>KLF14</i>	68.295	0.111

Abbreviations: Significant p-values are in bold.

E. DNA methylation age estimation through multi-tissues

Table S18: Univariate linear regression analysis of the 37 CpG sites in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* loci in 204 samples including blood from living and deceased individuals, teeth from living and deceased individuals and bone collected during autopsies.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 201)	CpG1	Chr6: 11044628	0.700	0.490	0.488	16.38	6.34 × 10⁻³¹
	CpG2	Chr6: 11044631	0.614	0.377	0.374	18.10	3.36 × 10⁻²²
	CpG3	Chr6: 11044634	0.653	0.427	0.424	17.37	7.77 × 10⁻²⁶
	CpG4	Chr6: 11044640	0.603	0.364	0.361	18.29	2.51 × 10⁻²¹
	CpG5	Chr6: 11044642	0.706	0.499	0.496	16.24	1.12 × 10⁻³¹
	CpG6	Chr6: 11044644	0.696	0.484	0.481	16.48	2.09 × 10⁻³⁰
	CpG7	Chr6: 11044647	0.595	0.354	0.351	18.44	1.27 × 10⁻²⁰
	CpG8	Chr6: 11044655	0.629	0.395	0.392	17.84	1.66 × 10⁻²³
	CpG9	Chr6: 11044661	0.575	0.331	0.328	18.76	4.17 × 10⁻¹⁹
<i>FHL2</i> (N = 203)	CpG1	Chr2:105399282	0.662	0.438	0.435	17.07	5.92 × 10⁻²⁷
	CpG2	Chr2: 105399288	0.605	0.366	0.363	18.12	1.10 × 10⁻²¹
	CpG3	Chr2: 105399291	0.614	0.377	0.374	17.97	1.99 × 10⁻²²
	CpG4	Chr2: 105399297	0.512	0.262	0.258	19.56	6.09 × 10⁻¹⁵
	CpG5	Chr2: 105399300	0.626	0.392	0.389	17.76	1.84 × 10⁻²³
	CpG6	Chr2: 105399310	0.527	0.277	0.274	19.36	7.14 × 10⁻¹⁶
	CpG7	Chr2: 105399314	0.313	0.098	0.094	21.62	0.000005
	CpG8	Chr2: 105399316	0.387	0.150	0.145	20.99	1.20 × 10⁻⁸
	CpG9	Chr2: 105399323	0.079	0.006	0.001	22.70	0.264956
	CpG10	Chr2: 105399327	0.024	0.001	-0.004	22.76	0.738171
	CpG11	Chr2: 105399338	-0.460	0.211	0.207	20.22	5.23 × 10⁻¹²
	CpG12	Chr2: 105399340	-0.130	0.017	0.012	22.56	0.064529
<i>EDARADD</i> (N = 202)	CpG1	Chr1: 236394458	-0.030	0.001	-0.004	22.75	0.666693
	CpG2	Chr1: 236394441	-0.517	0.267	0.264	19.49	3.41 × 10⁻¹⁵
	CpG3	Chr1: 236394382	-0.682	0.465	0.463	16.64	5.27 × 10⁻²⁹
	CpG4	Chr1: 236394370	-0.403	0.163	0.159	20.83	2.61 × 10⁻⁹
<i>PDE4C</i> (N = 196)	CpG1	Chr19: 18233139	0.497	0.247	0.243	19.97	1.11 × 10⁻¹³
	CpG2	Chr19: 18233133	0.605	0.366	0.363	18.32	4.50 × 10⁻²¹
	CpG3	Chr19: 18233131	0.562	0.316	0.313	19.03	8.13 × 10⁻¹⁸
	CpG4	Chr19: 18233127	0.246	0.060	0.055	22.31	0.000506
	CpG5	Chr19: 18233105	0.456	0.208	0.204	20.49	1.69 × 10⁻¹¹
	CpG6	Chr19: 18233091	-0.141	0.020	0.015	22.79	0.048212
	CpG7	Chr19: 18233082	0.058	0.003	-0.002	22.98	0.419384
	CpG8	Chr19: 18233079	-0.135	0.018	0.013	22.80	0.058330
	CpG9	Chr19: 18233070	-0.298	0.089	0.084	21.97	0.000021
	CpG10	Chr19: 18233058	-0.042	0.002	-0.003	23.01	0.556208
	CpG11	Chr19: 18233048	-0.095	0.009	0.004	23.03	0.192072
	CpG12	Chr19: 18233042	-0.014	0.000	-0.005	23.13	0.843371

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S19: Univariate linear regression analysis of the 43 CpG sites of *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* loci in 133 samples including blood from deceased individuals, teeth from living and deceased individuals and bone collected during autopsies.

Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 130)	CpG1	Chr6: 11044628	0.516	0.266	0.260	14.34	3.42 × 10⁻¹⁰
	CpG2	Chr6: 11044631	0.508	0.258	0.253	14.41	6.63 × 10⁻¹⁰
	CpG3	Chr6: 11044634	0.439	0.192	0.186	15.04	1.78 × 10⁻⁷
	CpG4	Chr6: 11044640	0.447	0.200	0.194	14.97	9.56 × 10⁻⁸
	<u>CpG5</u>	Chr6: 11044642	0.563	0.317	0.312	13.84	3.09 × 10⁻¹²
	CpG6	Chr6: 11044644	0.524	0.274	0.269	14.26	1.61 × 10⁻¹⁰
	CpG7	Chr6: 11044647	0.425	0.181	0.174	15.15	4.57 × 10⁻⁷
	CpG8	Chr6: 11044655	0.488	0.238	0.232	14.61	3.82 × 10⁻⁹
	CpG9	Chr6: 11044661	0.478	0.229	0.223	14.70	8.56 × 10⁻⁹
<i>FHL2</i> (N = 132)	<u>CpG1</u>	Chr2:105399282	0.365	0.133	0.127	15.35	0.000017
	CpG2	Chr2:105399288	0.342	0.117	0.110	15.49	0.000060
	CpG3	Chr2:105399291	0.357	0.127	0.121	15.40	0.000027
	CpG4	Chr2:105399297	0.262	0.069	0.062	15.91	0.002368
	CpG5	Chr2:105399300	0.360	0.129	0.123	15.38	0.000023
	CpG6	Chr2:105399310	0.266	0.071	0.064	15.89	0.002044
	CpG7	Chr2:105399314	0.210	0.044	0.037	16.12	0.015669
	CpG8	Chr2:105399316	0.257	0.066	0.059	15.93	0.002951
	CpG9	Chr2:105399323	0.092	0.008	0.001	16.42	0.295906
	CpG10	Chr2:105399327	0.040	0.002	-0.006	16.47	0.652528
	CpG11	Chr2:105399338	-0.149	0.022	0.015	16.30	0.088475
	CpG12	Chr2:105399340	-0.033	0.001	0.007	16.48	0.704589
<i>EDARADD</i> (N = 131)	CpG1	Chr1:236394458	-0.034	0.001	-0.007	16.51	0.66260
	CpG2	Chr1:236394441	-0.254	0.065	0.057	15.98	0.003365
	<u>CpG3</u>	Chr1:236394382	-0.416	0.173	0.166	15.03	7.90 × 10⁻⁷
	CpG4	Chr1:236394370	-0.305	0.093	0.086	15.74	0.000395
<i>PDE4C</i> (N = 125)	CpG1	Chr19:18233139	0.362	0.131	0.124	15.52	0.000031
	<u>CpG2</u>	Chr19:18233133	0.470	0.221	0.214	14.70	2.87 × 10⁻⁸
	CpG3	Chr19:18233131	0.419	0.176	0.169	15.12	0.000001
	CpG4	Chr19:18233127	0.307	0.094	0.087	15.85	0.000465
	CpG5	Chr19:18233105	0.314	0.099	0.091	15.81	0.000340
	CpG6	Chr19:18233091	0.063	0.004	-0.004	16.62	0.485909
	CpG7	Chr19:18233082	0.137	0.019	0.011	16.49	0.125778
	CpG8	Chr19:18233079	0.042	0.002	-0.006	16.64	0.642685
	CpG9	Chr19:18233070	-0.115	0.013	0.005	16.54	0.200083
	CpG10	Chr19:18233058	0.081	0.006	-0.002	16.10	0.373614
	CpG11	Chr19:18233048	-0.015	0.000	-0.008	16.55	0.872788
	CpG12	Chr19:18233042	0.032	0.001	-0.007	16.54	0.723825
<i>C1orf132</i> (N = 126)	<u>CpG1</u>	Chr1:207823681	-0.455	0.207	0.200	14.67	7.04 × 10⁻⁸
	CpG2	Chr1:207823675	-0.311	0.097	0.089	15.65	0.000356
	CpG3	Chr1:207823672	-0.197	0.039	0.031	16.15	0.025508
	CpG4	Chr1:207823660	-0.195	0.038	0.030	16.17	0.028147
	CpG5	Chr1:207823657	-0.335	0.112	0.105	15.47	0.000133
	CpG6	Chr1:207823637	-0.065	0.004	-0.004	16.46	0.468501

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Table S20: Univariate linear regression analysis of the nine CpG sites in gene *ELOVL2* in blood samples from living and deceased individuals, bone samples collected during autopsies, buccal swabs from living individuals and tooth samples from living and deceased individuals.

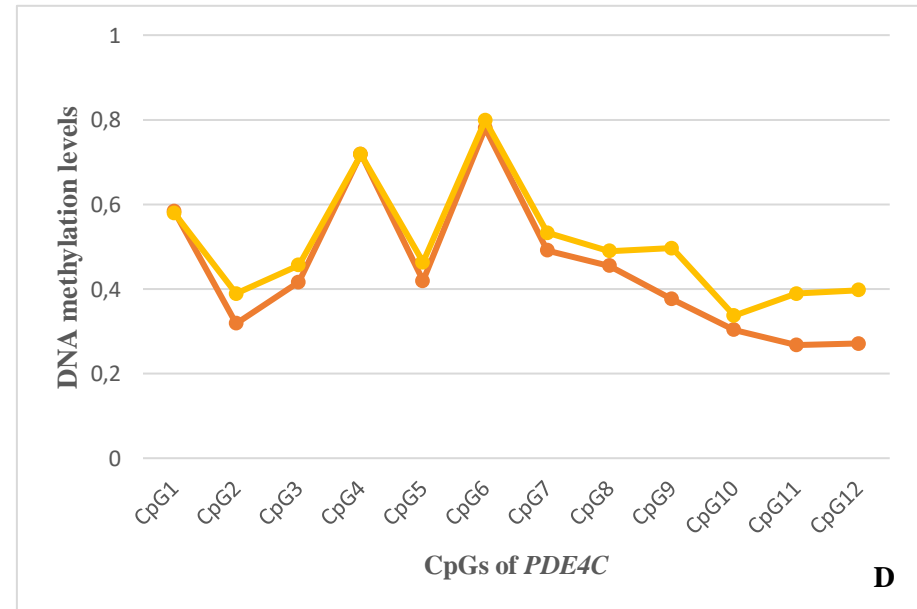
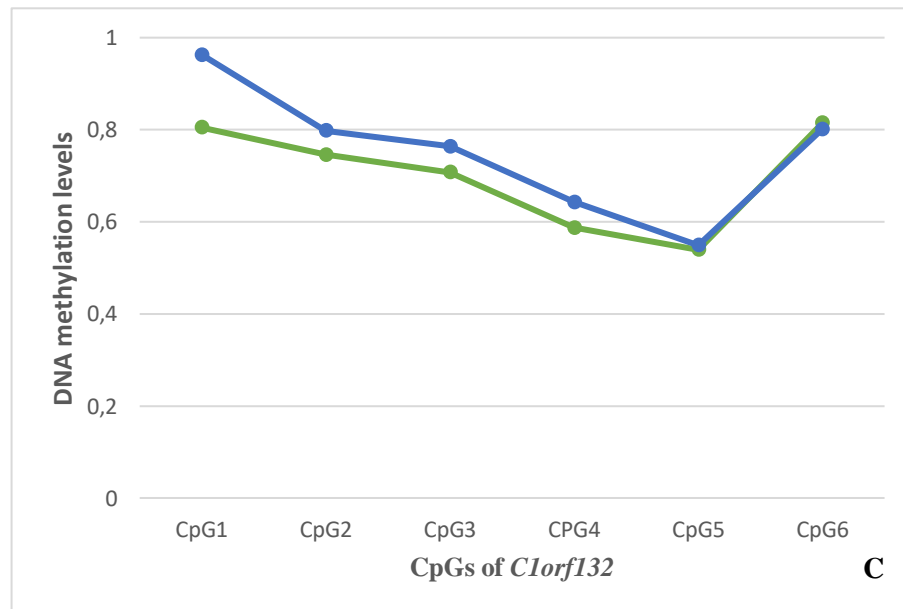
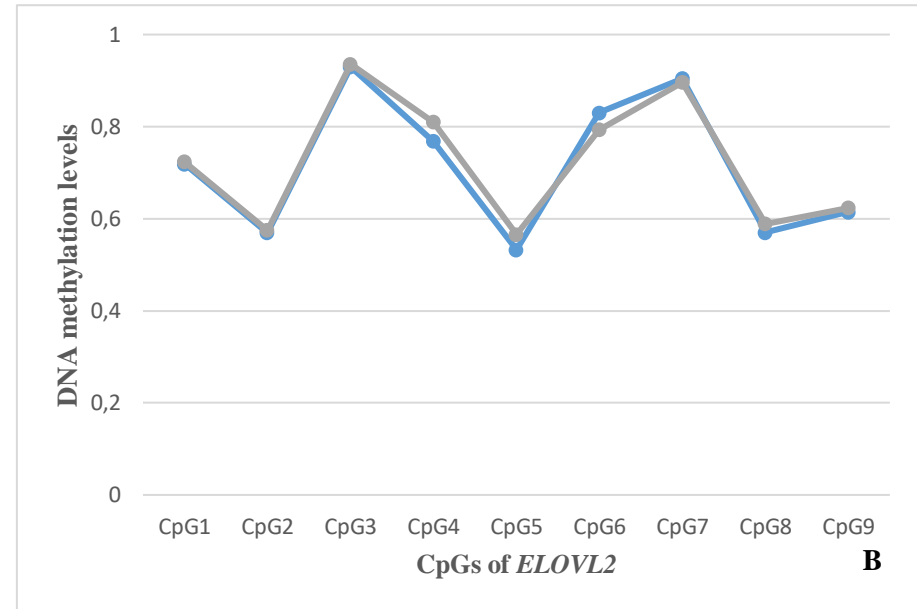
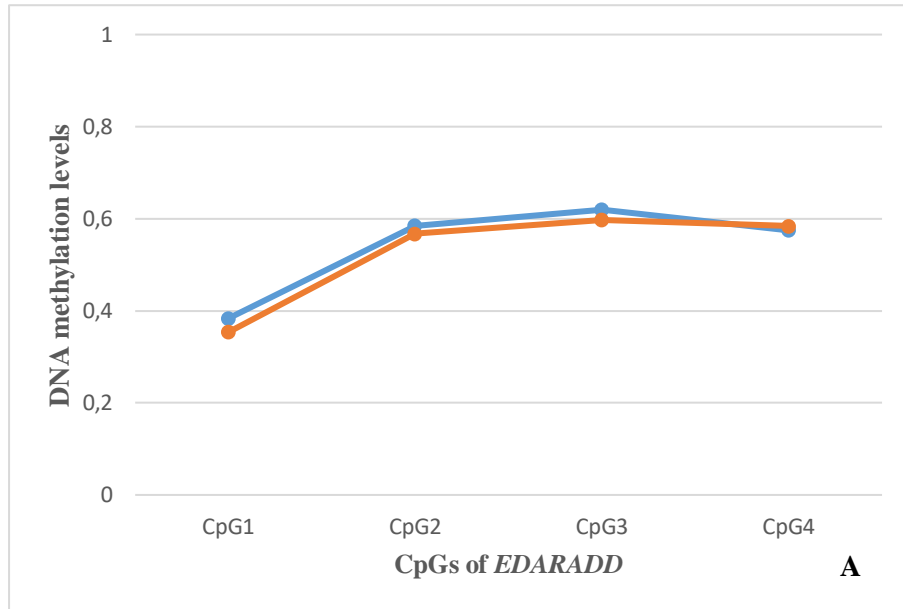
Gene	CpG site	Location	R	R ²	Corrected R ²	SE	P-value
<i>ELOVL2</i> (N = 224)	CpG1	Chr6: 11044628	0.712	0.507	0.505	17.06	5.58 × 10⁻³⁶
	CpG2	Chr6: 11044631	0.626	0.392	0.389	18.97	9.29 × 10⁻²⁶
	CpG3	Chr6: 11044634	0.678	0.460	0.458	17.86	1.50 × 10⁻³¹
	CpG4	Chr6: 11044640	0.634	0.402	0.399	18.80	1.50 × 10⁻²⁶
	<u>CpG5</u>	Chr6: 11044642	0.736	0.541	0.539	16.46	2.07 × 10⁻³⁹
	CpG6	Chr6: 11044644	0.708	0.501	0.499	17.16	2.08 × 10⁻³⁵
	CpG7	Chr6: 11044647	0.618	0.381	0.379	19.12	6.12 × 10⁻²⁵
	CpG8	Chr6: 11044655	0.661	0.437	0.434	18.25	1.80 × 10⁻²⁹
	CpG9	Chr6: 11044661	0.582	0.338	0.335	19.77	1.16 × 10⁻²¹

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error. Significant p-values are in bold. The strongest age-associated CpG site is in bold and underlined. Genomic positions were based on the GRCh38/hg38 assembly.

Annex VI
Supplementary Material Figures

A. DNA methylation age estimation in blood samples

Annex VI. Supplementary Material Figures



DNA methylation as an age predictor in living and deceased individuals

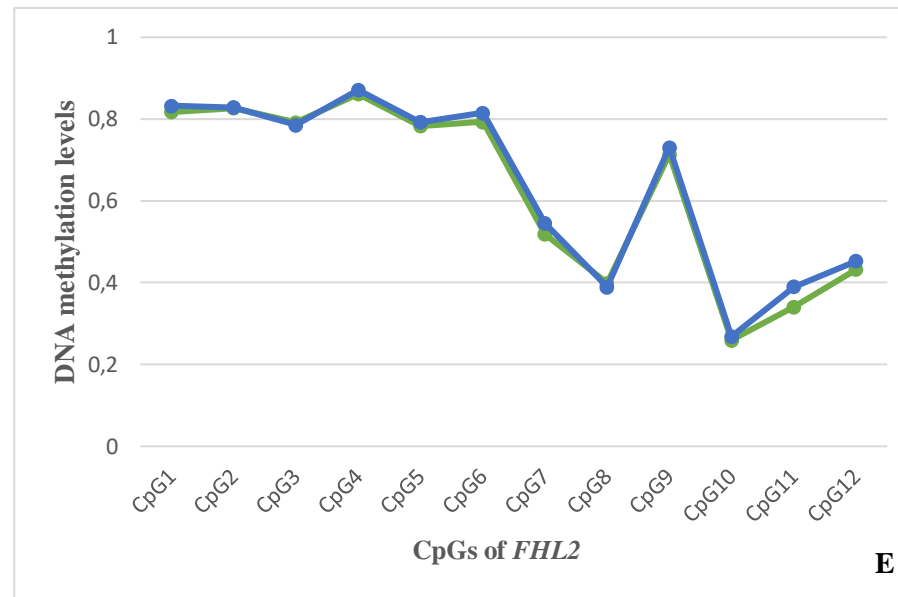


Figure S1: Reproducibility of direct bisulfite Sanger sequencing. Each plot represents DNAm levels of each gene obtained from one sample (one individual) that was run in duplicate analysis. A) DNAm levels of the four CpGs located at *EDARADD* gene in a blood sample from a male, aged 77 years old; B) DNAm levels of the nine CpGs located at *ELOVL2* gene in a blood sample from a male, aged 55 years old; C) DNAm levels of the six CpGs located at *C1orf132* gene in a blood sample from a male, aged 28 years old; D) DNAm levels of the 12 CpGs located at *PDE4C* gene in a blood sample from a female, aged 2 years old; E) DNAm levels of the 12 CpGs located at *FHL2* gene in a blood sample from a female, aged 76 years old. Colors represent two separate analysis of the same individual (PCR amplification and Sanger sequencing). The concordance between the two colors within each plot showed the reproducibility of direct bisulfite Sanger sequencing method.

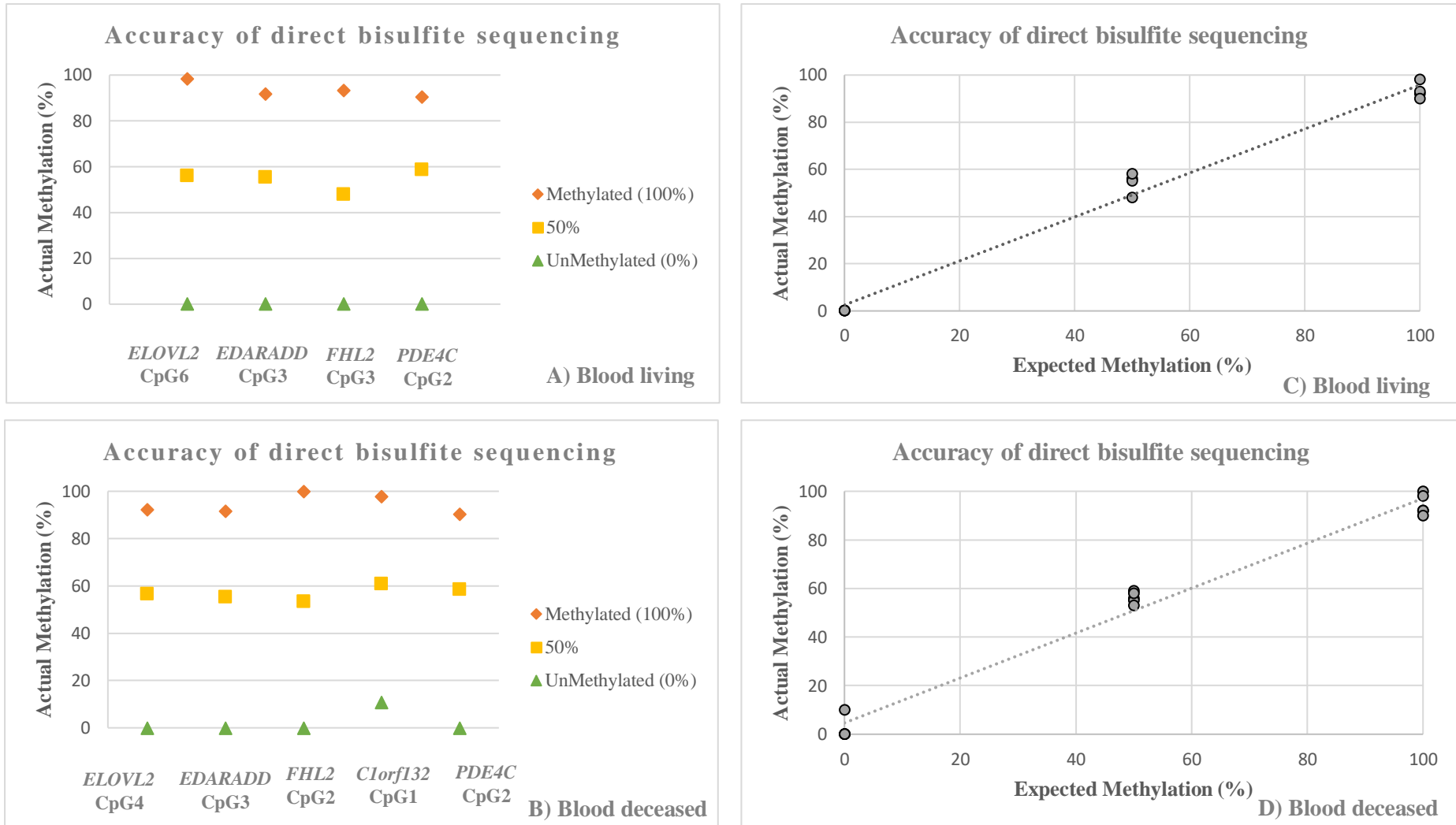


Figure S2: Actual methylation (0–100%) obtained for the best-selected CpGs in blood samples from living (A) and deceased individuals (B). Actual methylation *versus* expected methylation of known quantities of methylated to unmethylated DNA standards in blood samples from living (C) and deceased individuals (D).

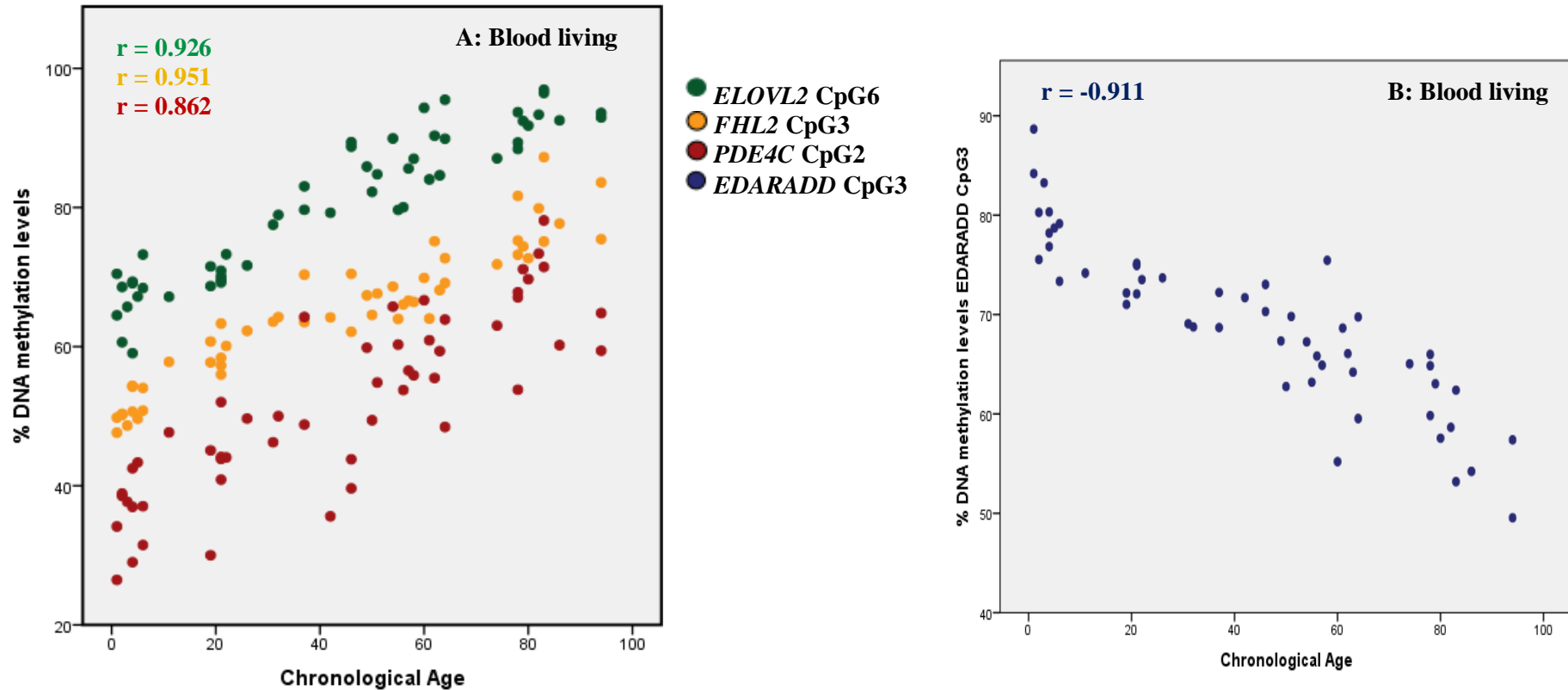
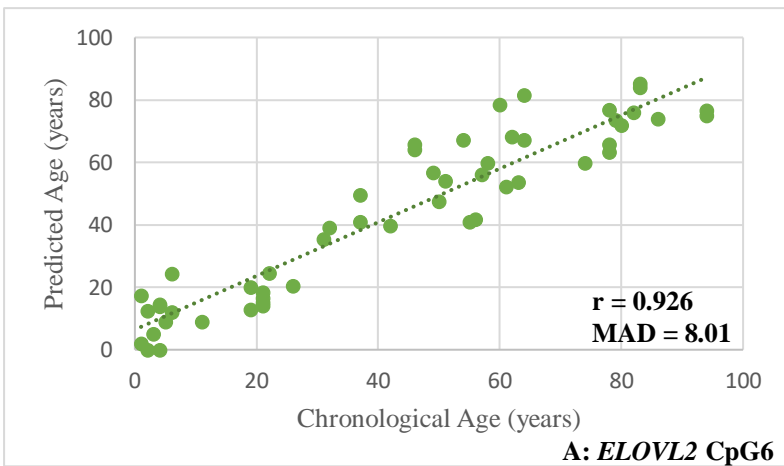
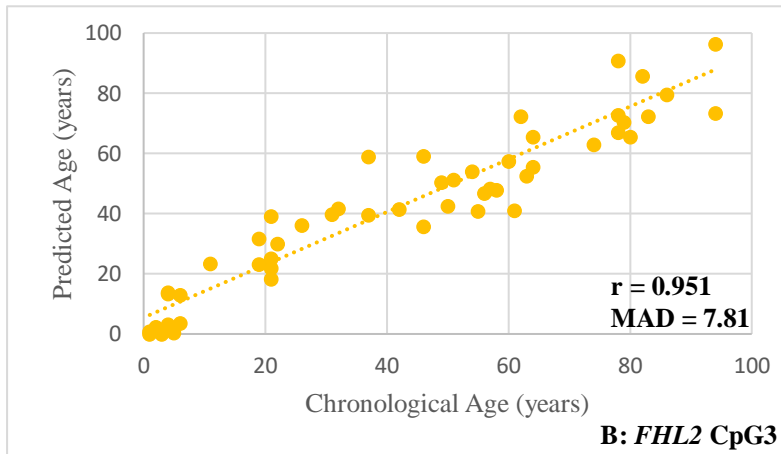


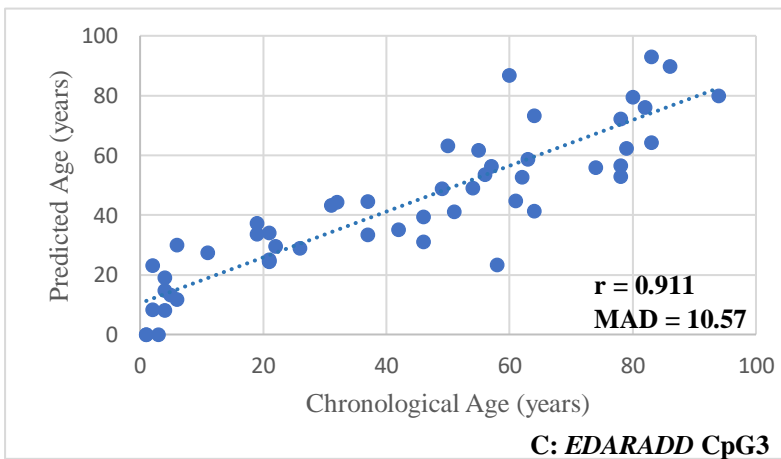
Figure S3: Correlations between DNAm levels and chronological age in 53 blood samples from living individuals obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *ELOVL2* CpG6 (green), *FHL2* CpG3 (yellow) and *PDE4C* CpG2 (dark red) markers; B) negative correlation for *EDARADD* CpG3 marker (blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



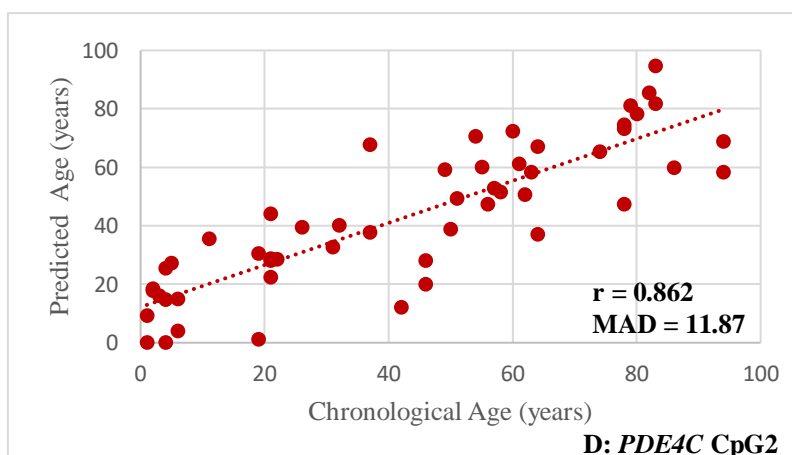
A) $Predicted\ Age = (-163.384) + 256.420 \times DNAm\ level\ ELOVL2\ CpG6.$



B) $Predicted\ Age = (-139.712) + 282.380 \times DNAm\ level\ FHL2\ CpG3.$



C) $Predicted\ Age = 259.544 - 313.003 \times DNAm\ level\ EDARADD\ CpG3.$



$$D) \text{ Predicted Age} = (-57.040) + 194.157 \times \text{DNAm level } PDE4C \text{ CpG2.}$$

Figure S4: Plots with predicted age (years) *versus* chronological age (years) of the 53 living individuals using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient are plotted in each chart. A) Predicted age of the 53 living individuals based on methylation levels of the *ELOVL2* CpG6. For two individuals aged 2 and 4 years old, negative prediction values were obtained and were set at 0; B) Predicted age of the 53 living individuals based on methylation levels of the *FHL2* CpG3. For two individuals aged 1 and 3 years old, negative prediction values were obtained and were set at 0; C) Predicted age of the 53 living individuals based on methylation levels of the *EDARADD* CpG3. For three individuals aged 1 and 3 years old, negative prediction values were obtained and were set at 0; D) Predicted age of the 53 living individuals based on methylation levels of the *PDE4C* CpG2. For two individuals aged 1 and 4 years old, negative prediction values were obtained and were set at 0.

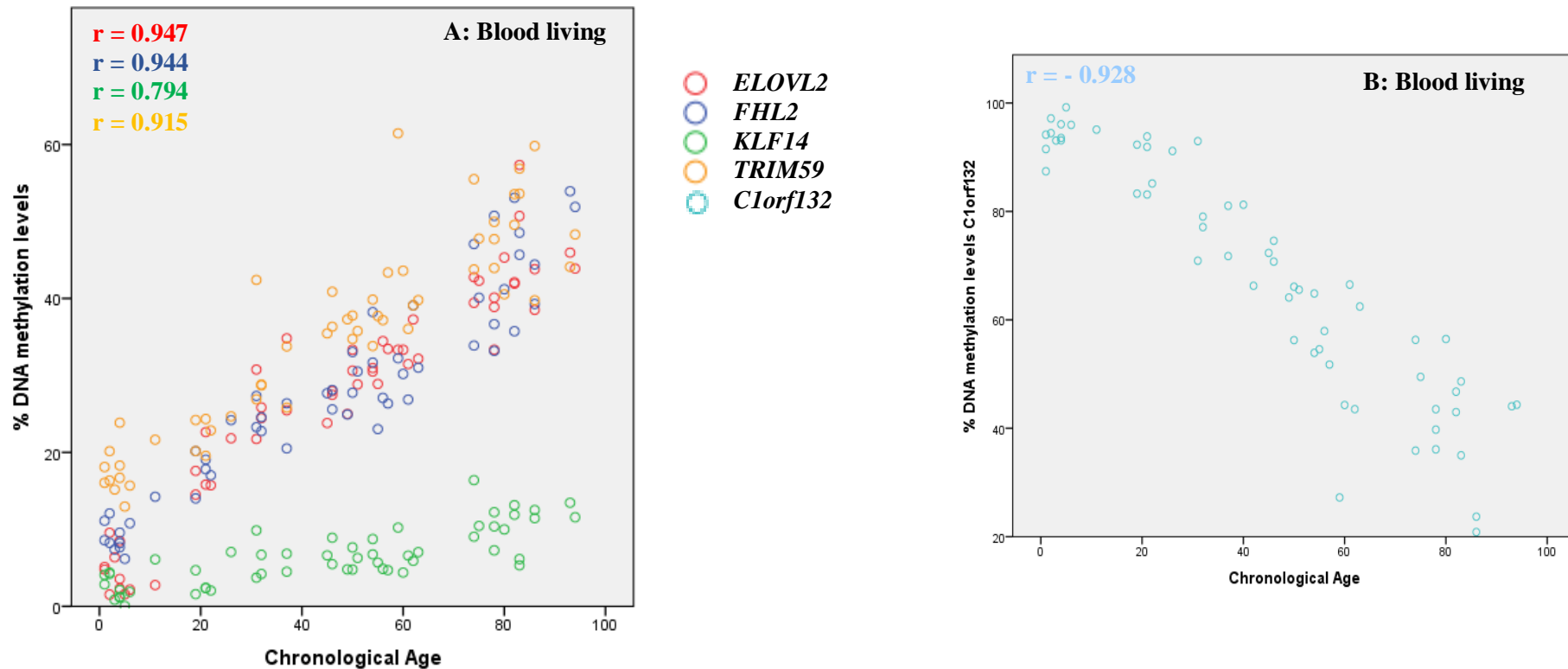
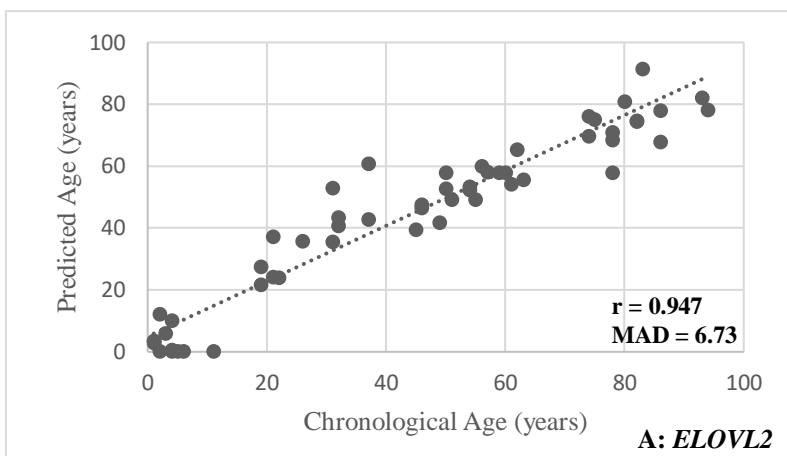
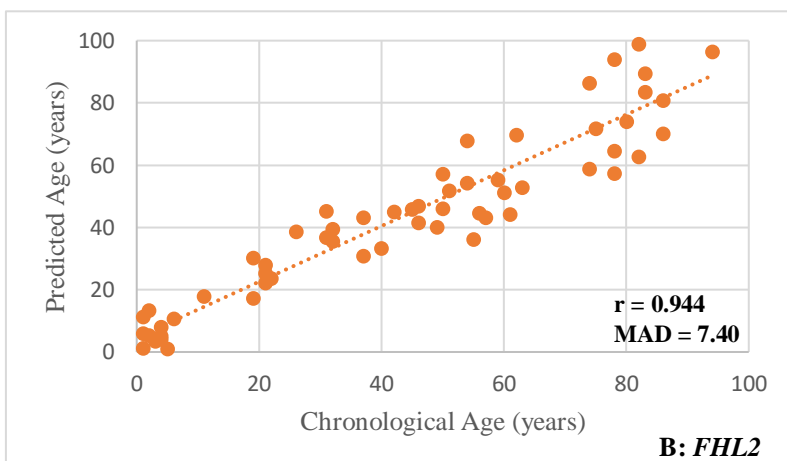


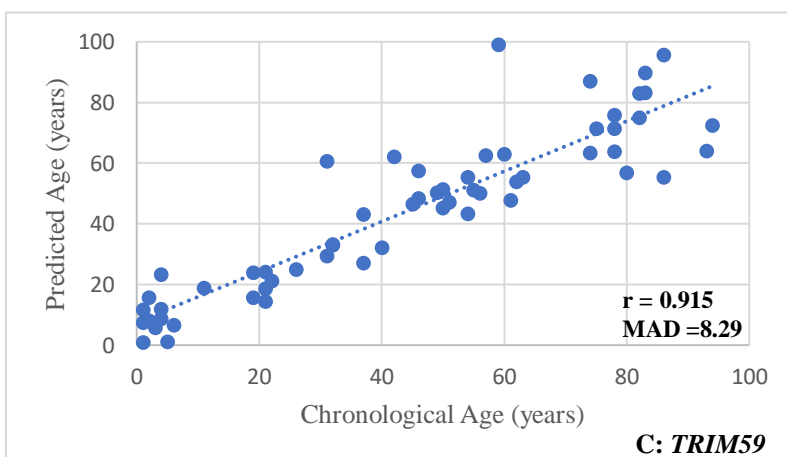
Figure S5: Correlations between DNAm levels and chronological age in 59 blood samples from living individuals obtained through SNaPshot methodology. A) Positive correlation between methylation levels and chronological age for CpG sites in *ELOVL2* (red), *FHL2* (dark blue), *KLF14* (green) and *TRIM59* (yellow) genes; B) negative correlation for *C1orf132* locus (light blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



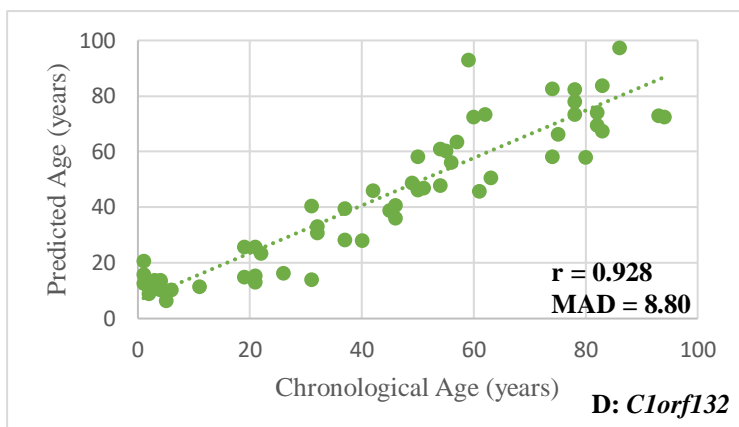
A) $Predicted\ Age = (-6.450) + 192.606 \times DNAm\ level\ ELOVL2.$



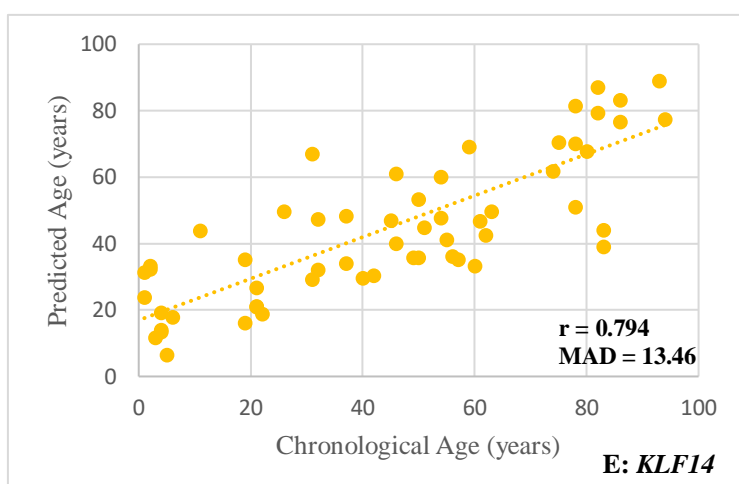
B) $Predicted\ Age = (-11.948) + 208.845 \times DNAm\ level\ FHL2.$



C) $Predicted\ Age = (-24.968) + 201.786 \times DNAm\ level\ TRIM59.$



$$D) \text{ Predicted Age} = 125.756 - 120.248 \times \text{DNAm level } C1orf132.$$



$$E) \text{ Predicted Age} = 6.352 + 613.667 \times \text{DNAm level } KLF14.$$

Figure S6: Plots with predicted age (years) versus chronological age (years) of the 59 living individuals using simple linear regression models developed with the five CpGs in each gene through the multiplex methylation SNaPshot methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 56 living individuals based on methylation levels of the CpG from *ELOVL2* gene. For the five individuals aged 2, 4, 5, 6 and 11 years old, negative prediction values were obtained and were set at 0; B) Predicted age of the 59 living individuals based on methylation levels of the CpG from *FHL2* gene; C) Predicted age of the 58 living individuals based on methylation levels of the CpG from *KLF14* gene; D) Predicted age of the 59 living individuals based on methylation levels of the CpG from *C1orf132* gene; E) Predicted age of the 59 living individuals based on methylation levels of the CpG from *TRIM59* gene.

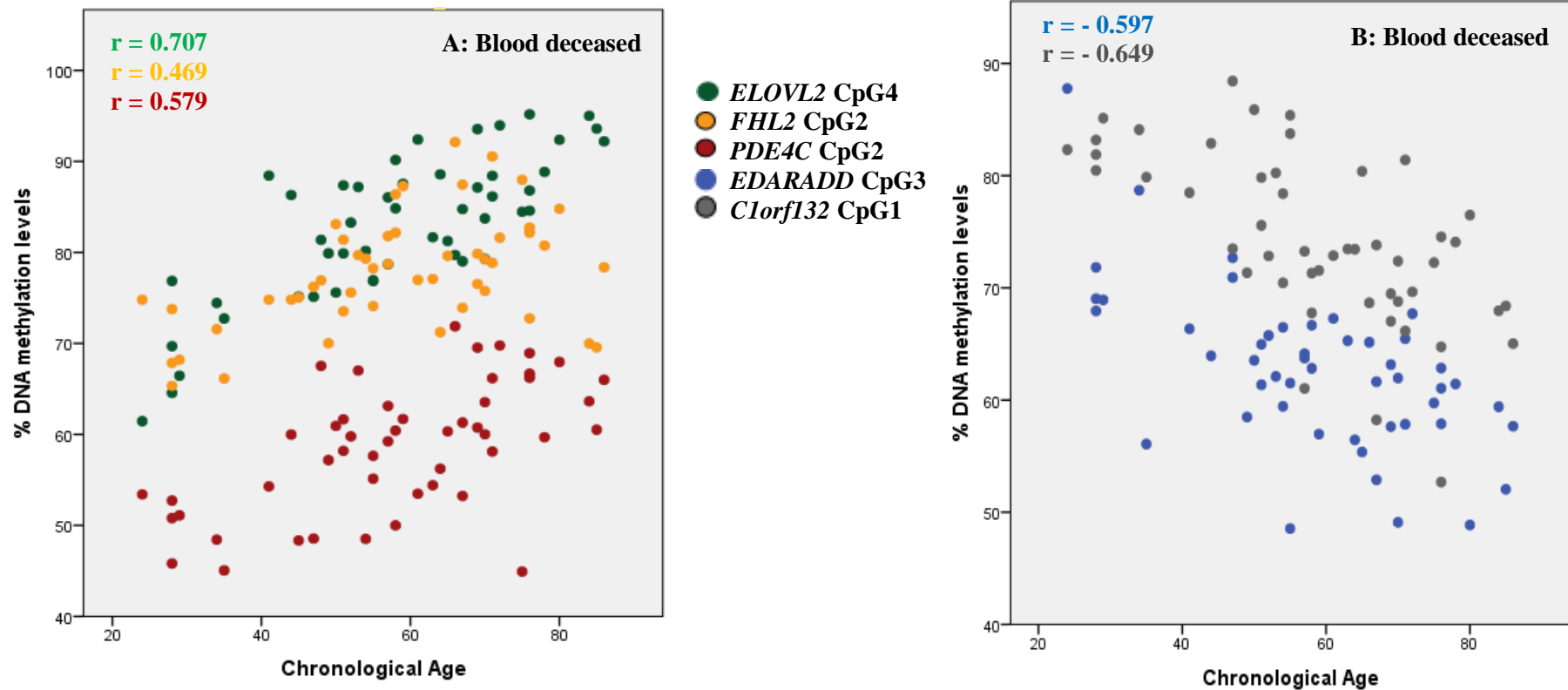
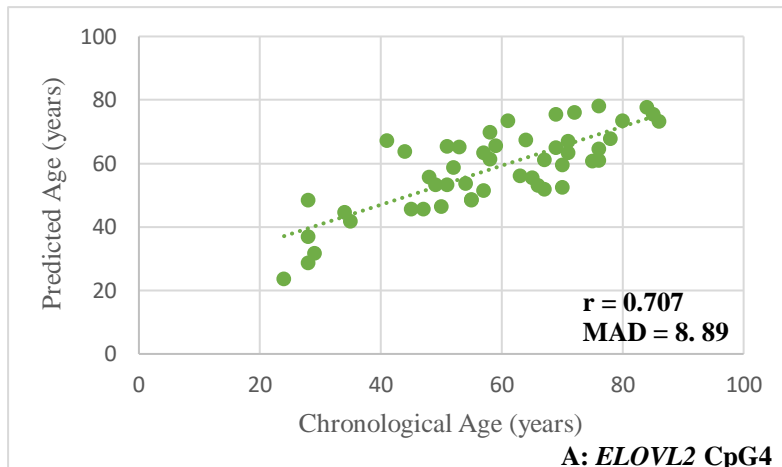
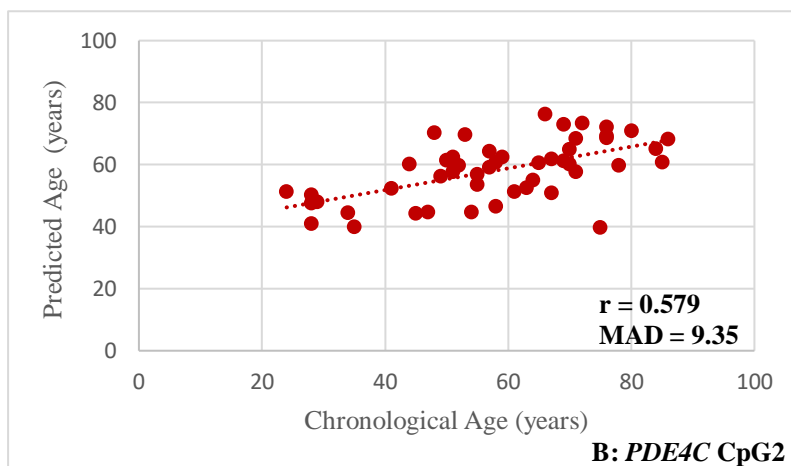


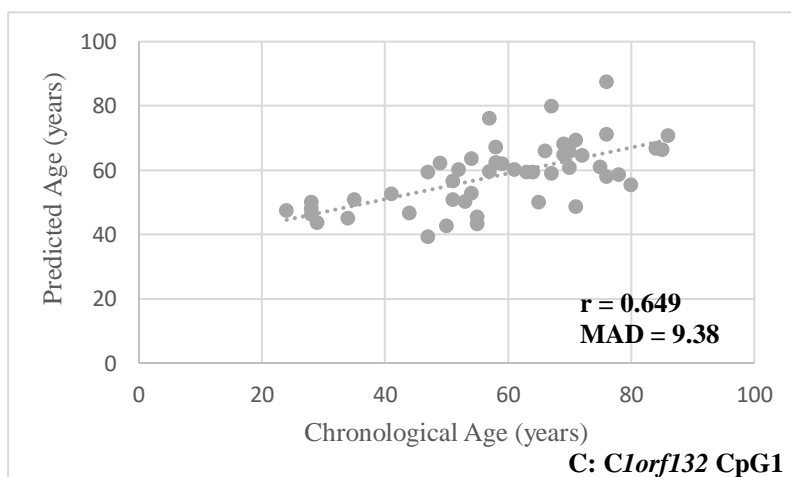
Figure S7: Correlations between DNAm levels and chronological age in 51 blood samples from deceased individuals obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *ELOVL2* CpG4 (green), *PDE4C* CpG2 (dark red), *FHL2* CpG2 (yellow) markers; B) negative correlation between methylation levels and chronological age for *C1orf132* CpG1 (gray) and *EDARADD* CpG3 (blue) markers. The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



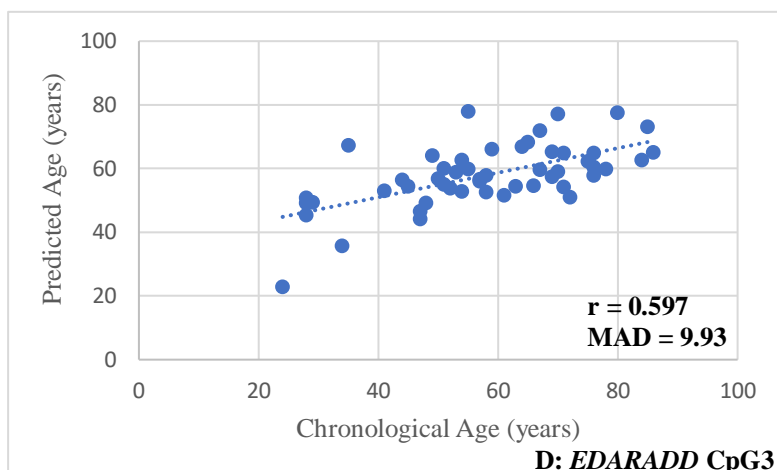
A) $Predicted\ Age = (-75.526) + 161.385 \times DNAm\ level\ ELOVL2\ CpG4.$



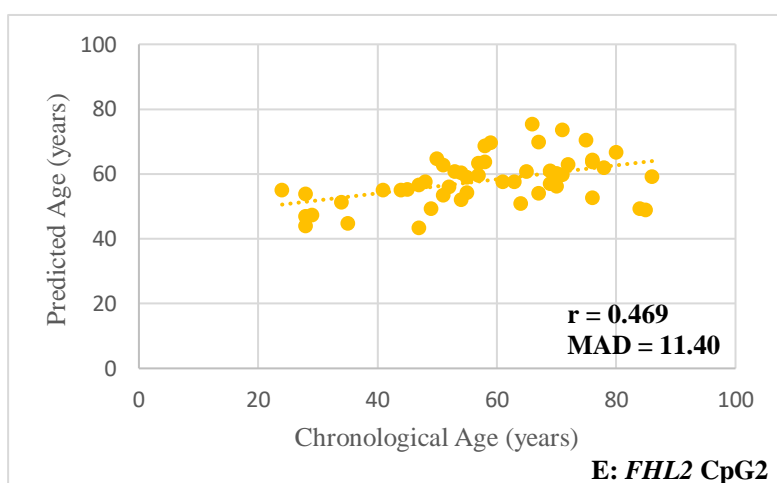
B) $Predicted\ Age = (-21.136) + 135.430 \times DNAm\ level\ PDE4C\ CpG2.$



C) $Predicted\ Age = 158.779 - 135.171 \times DNAm\ level\ Clorf132\ CpG1.$



D) $Predicted\ Age = 145.920 - 140.182 \times DNAm\ level\ EDARADD\ CpG3.$



E) $Predicted\ Age = (-33.136) + 117.698 \times DNAm\ level\ FHL2\ CpG2.$

Figure S8: Plots with predicted age (years) *versus* chronological age (years) of the 51 deceased individuals using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 49 deceased individuals based on methylation levels of the *ELOVL2* CpG4; B) Predicted age of the 49 deceased individuals based on methylation levels of the *PDE4C* CpG2; C) Predicted age of the 49 deceased individuals based on methylation levels of the *C1orf132* CpG1; D) Predicted age of the 51 deceased individuals based on methylation levels of the *EDARADD* CpG3; E) Predicted age of the 51 deceased individuals based on methylation levels of the *FHL2* CpG2.

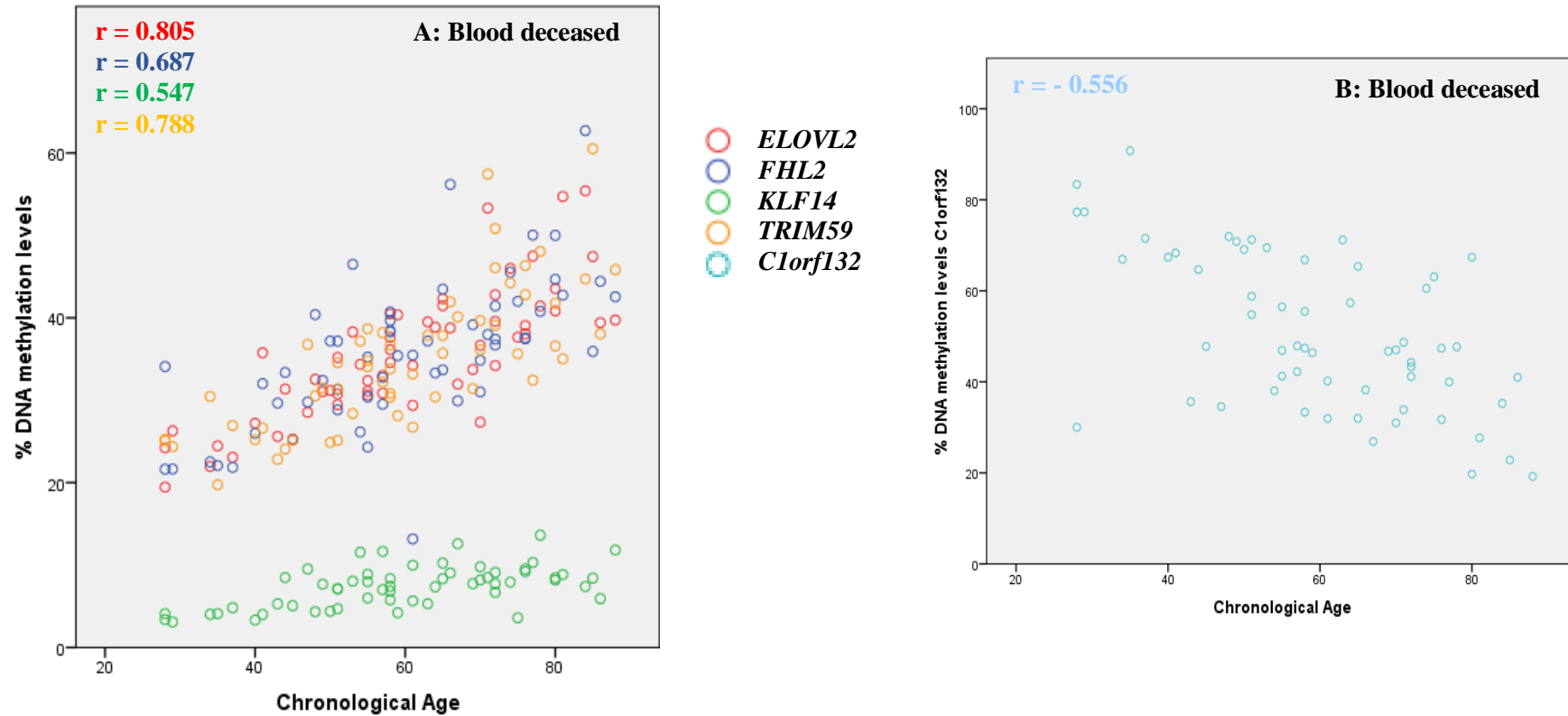
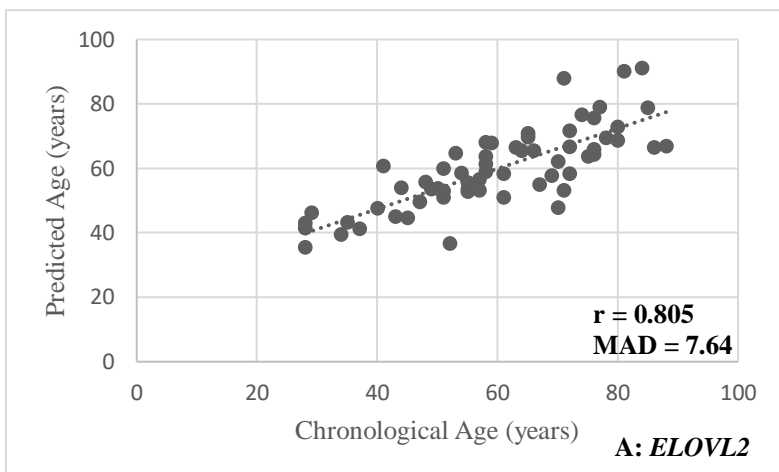
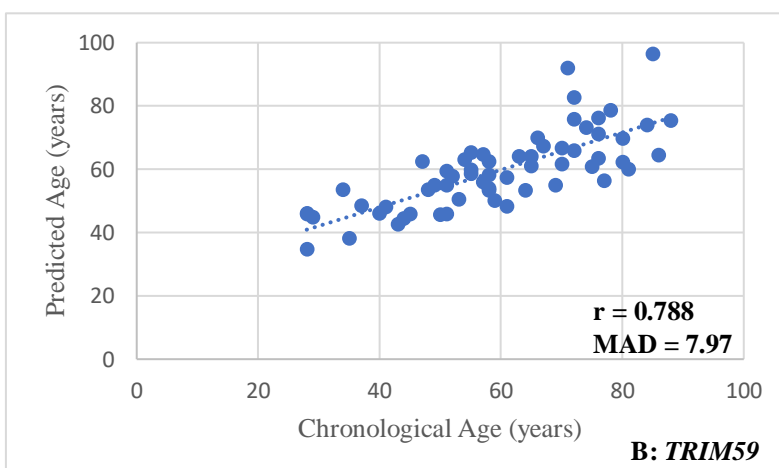


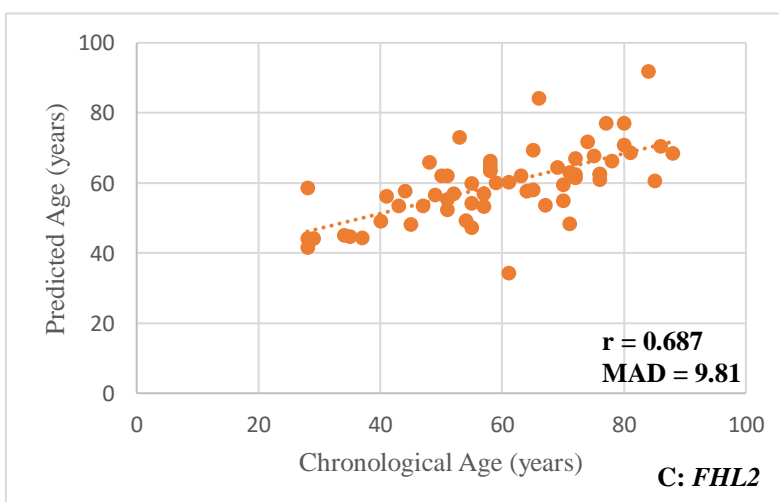
Figure S9: Correlations between DNAm levels and chronological age in 62 blood samples from deceased individuals obtained through SNaPshot methodology. A) Positive correlation between methylation levels and chronological age for CpG sites in *ELOVL2* (red), *FHL2* (dark blue), *KLF14* (green) and *TRIM59* (yellow) genes; B) negative correlation between methylation levels and chronological age for *C1orf132* locus (light blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



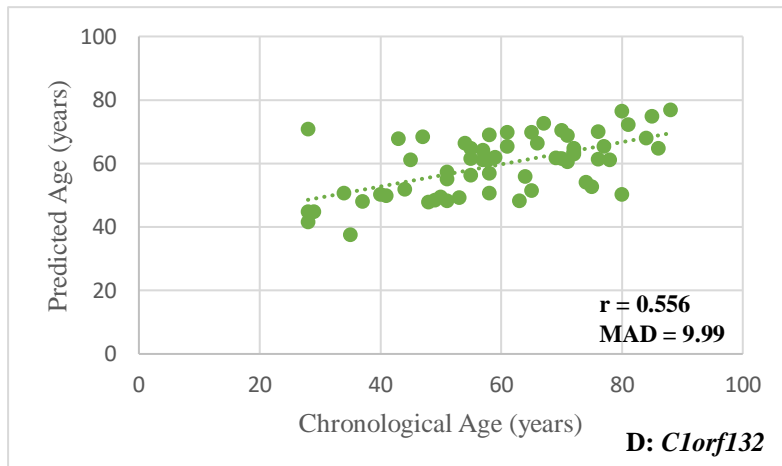
A) $Predicted\ Age = 5.446 + 154.579 \times DNAm\ level\ ELOVL2.$



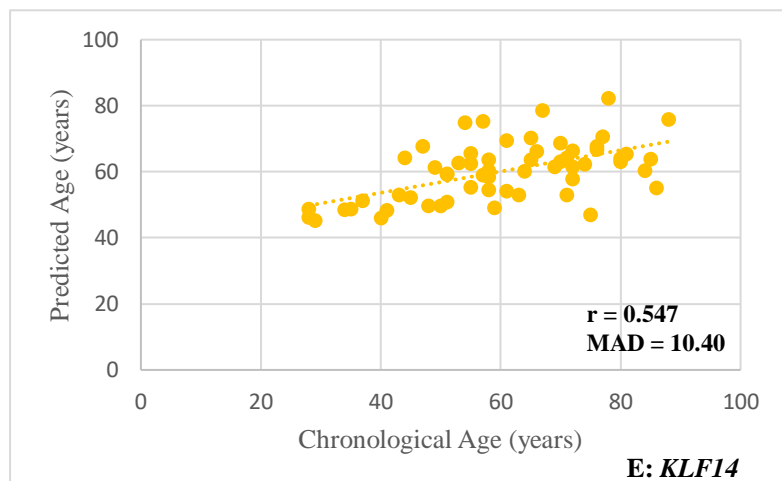
B) $Predicted\ Age = 9.947 + 142.829 \times DNAm\ level\ TRIM59.$



C) $Predicted\ Age = 19.012 + 116.043 \times DNAm\ level\ FHL2.$



$$D) \text{ Predicted Age} = 87.322 - 54.989 \times \text{DNAm level } C1orf132.$$



$$E) \text{ Predicted Age} = 34.305 + 351.694 \times \text{DNAm level } KLF14.$$

Figure S10: Plots with predicted age (years) *versus* chronological age (years) of the 62 deceased individuals using simple linear regression models developed with the five CpGs in each gene through the multiplex methylation SNaPshot methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 62 deceased individuals based on methylation levels of the CpG from *ELOVL2* gene; B) Predicted age of the 62 deceased individuals based on methylation levels of the CpG from *FHL2* gene; C) Predicted age of the 59 deceased individuals based on methylation levels of the CpG from *KLF14* gene; D) Predicted age of the 60 deceased individuals based on methylation levels of the CpG from *C1orf132* gene; E) Predicted age of the 61 deceased individuals based on methylation levels of the CpG from *TRIM59* gene.

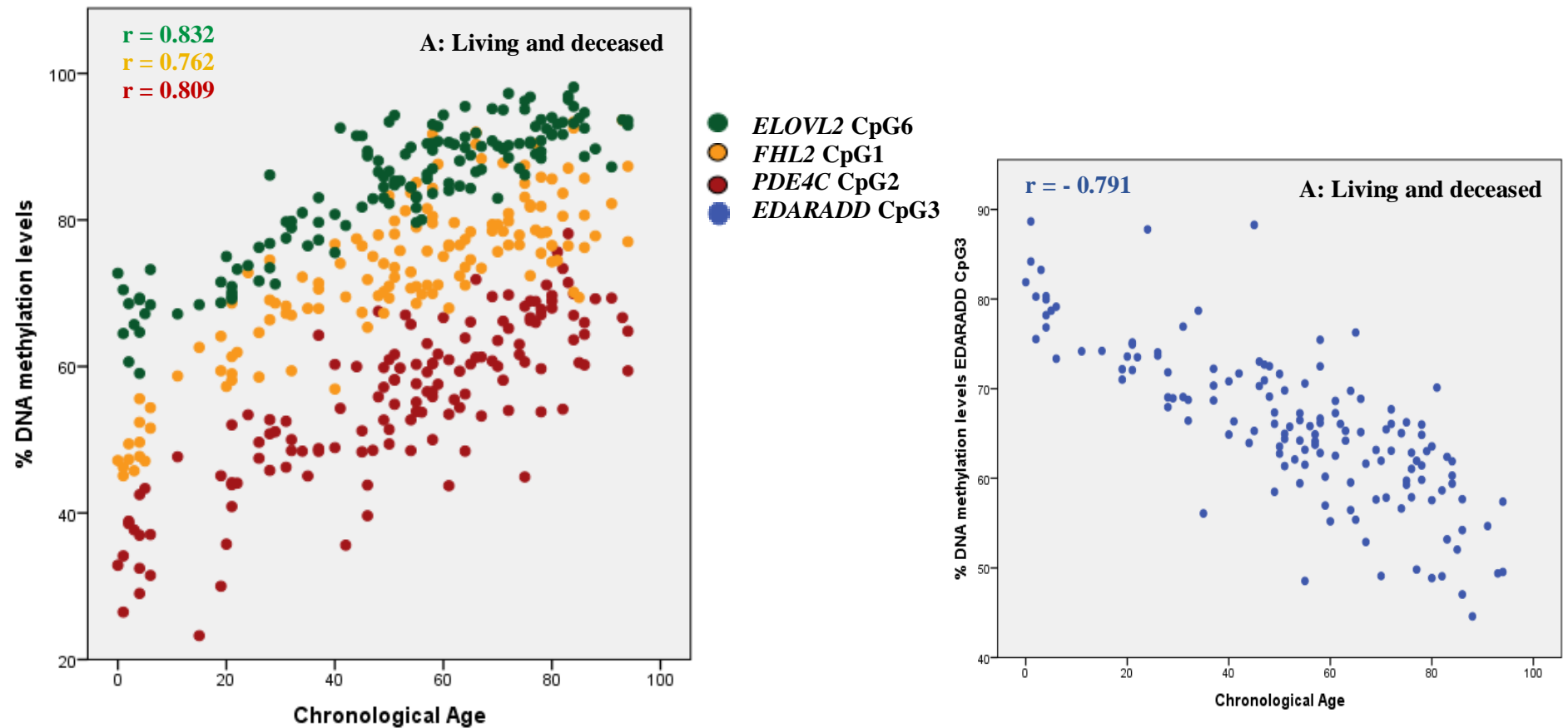
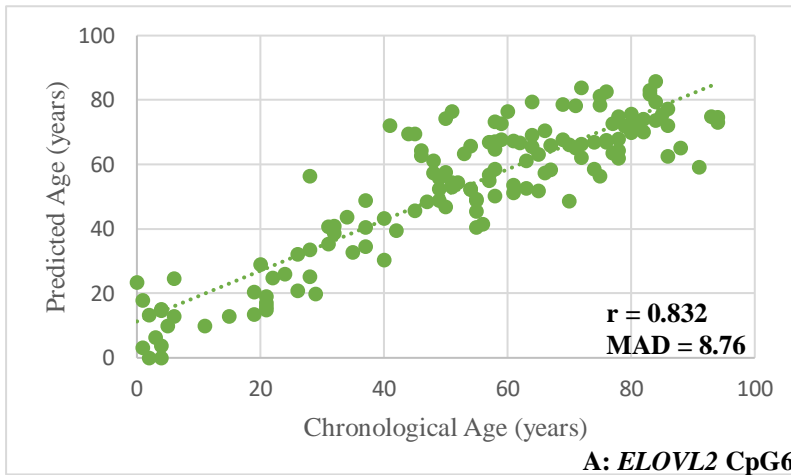
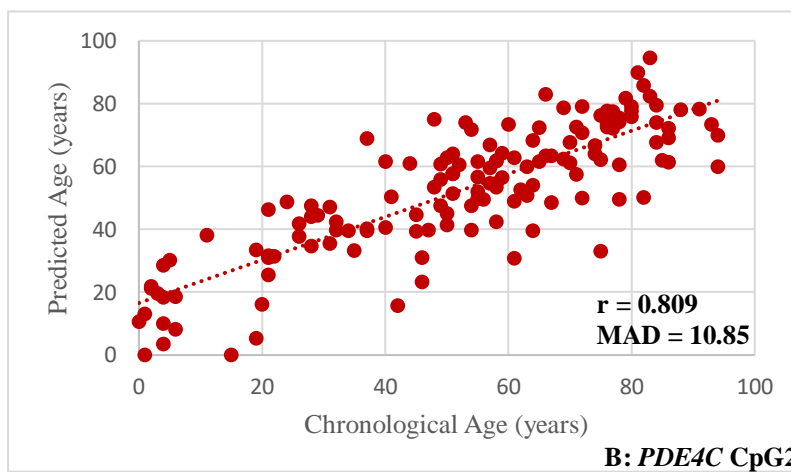


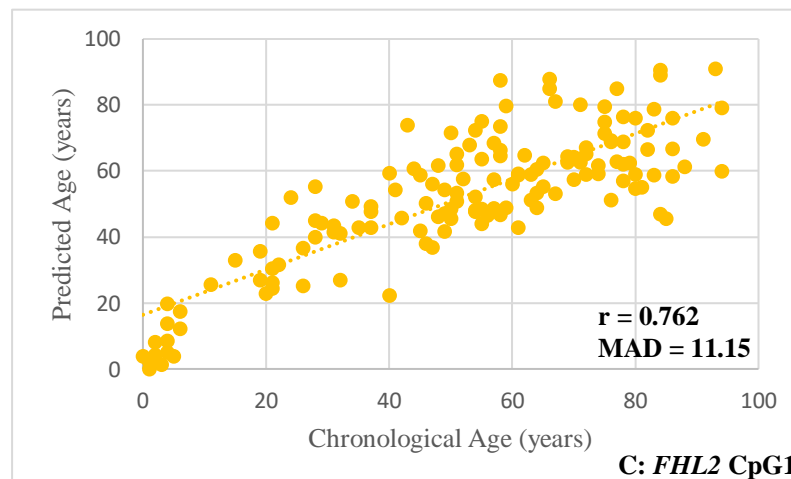
Figure S11: Correlations between DNAm levels and chronological age in 144 blood samples from living and deceased individuals obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and age for *ELOVL2* CpG6 (green), *FHL2* CpG1 (yellow) and *PDE4C* CpG2 (dark red) markers; B) negative correlation between methylation levels and chronological age for *EDARADD* CpG3 (blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



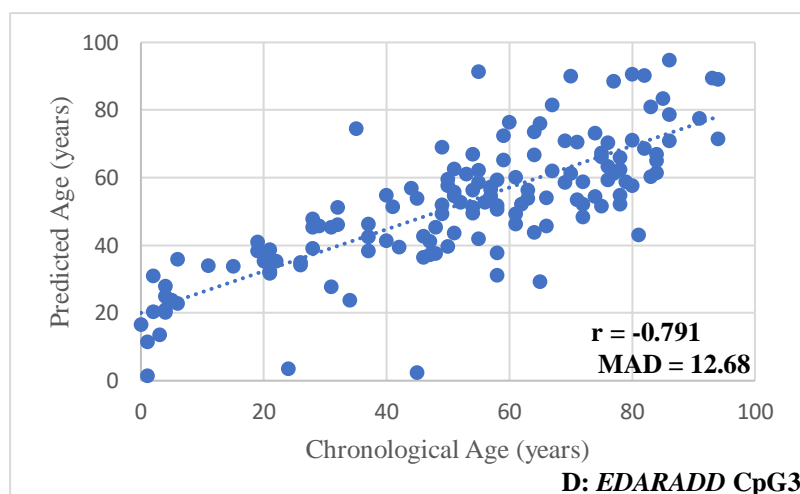
A) $Predicted\ Age = (-155.223) + 245.639 \times DNAm\ level\ ELOVL2\ CpG6.$



B) $Predicted\ Age = (-50.225) + 185.417 \times DNAm\ level\ PDE4C\ CpG2.$



C) $Predicted\ Age = (-84.436) + 187.461 \times DNAm\ level\ FHL2\ CpG1.$



D) $Predicted\ Age = 200.012 - 223.896 \times DNAm\ level\ EDARADD\ CpG3.$

Figure S12: Plots with predicted age (years) *versus* chronological age (years) of the 144 living and deceased individuals using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 141 living and deceased individuals based on methylation levels of the *ELOVL2* CpG6; For two individuals aged 2 and 4 years old, negative prediction values were obtained and were set at 0; B) Predicted age of the 141 living and deceased individuals based on methylation levels of the *PDE4C* CpG2. For two individuals aged 1 and 15 years old, negative prediction values were obtained and were set at 0; C) Predicted age of the 144 living and deceased individuals based on methylation levels of the *FHL2* CpG1; D) Predicted age of the 143 living and deceased individuals based on methylation levels of the *EDARADD* CpG3.

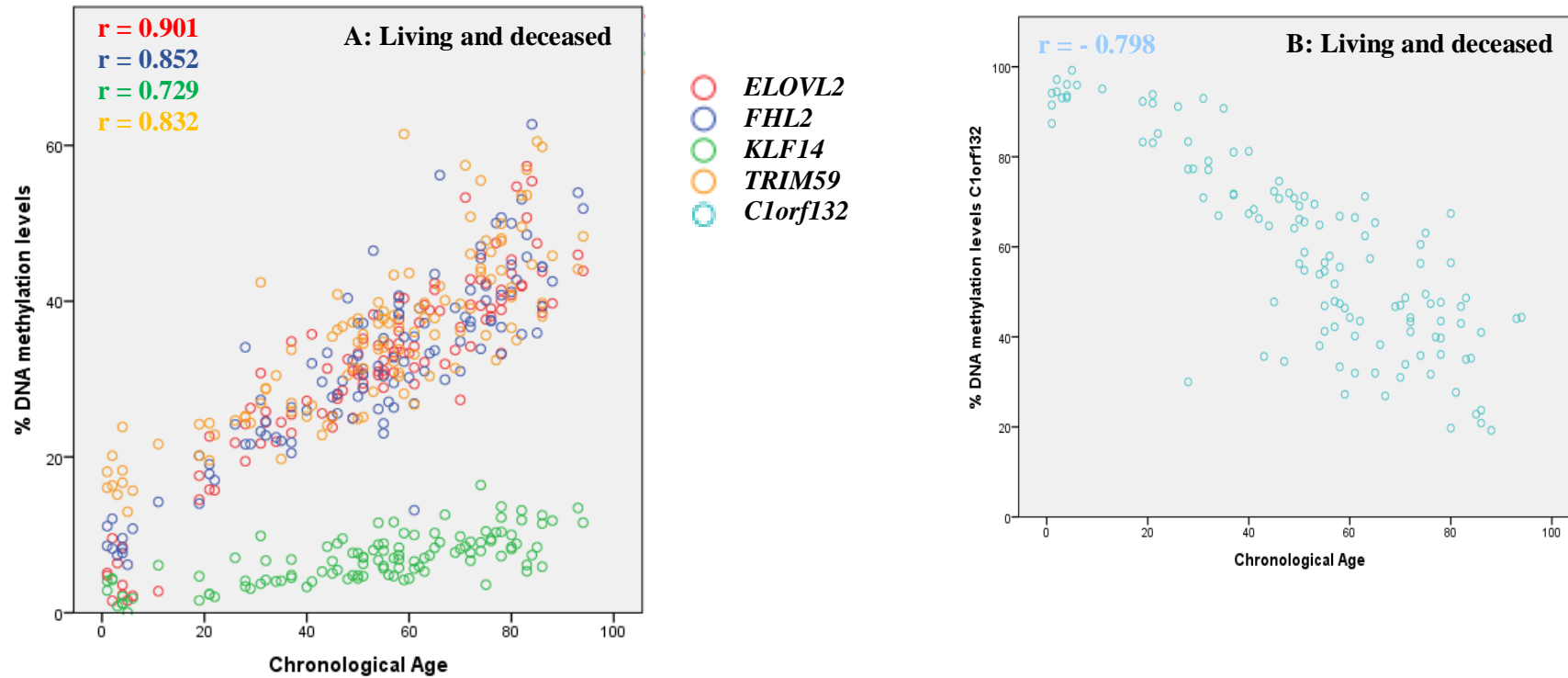
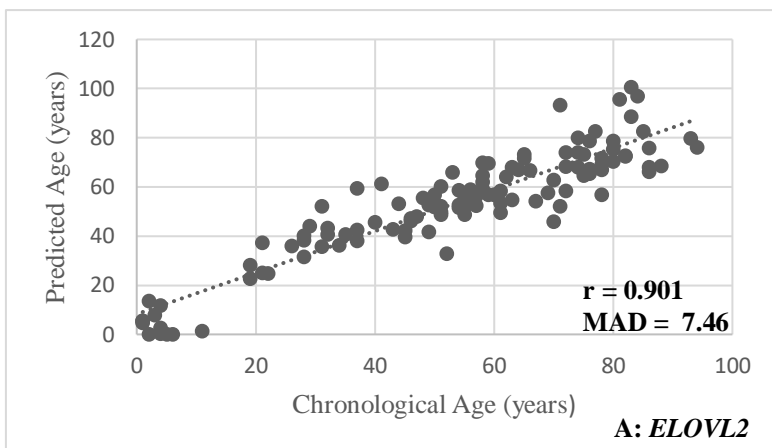
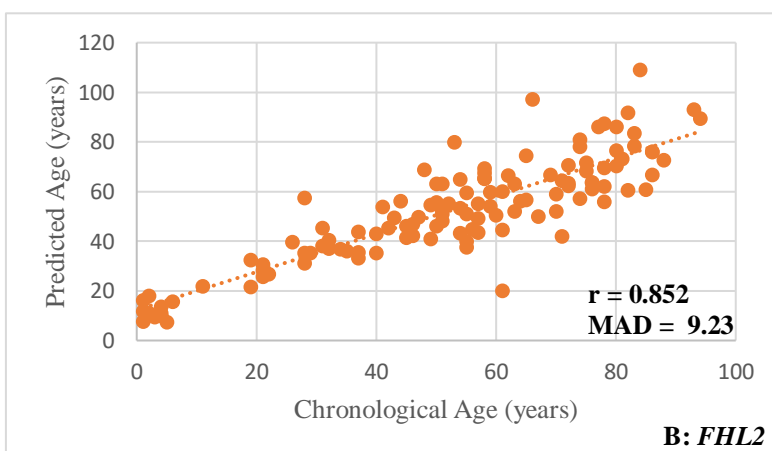


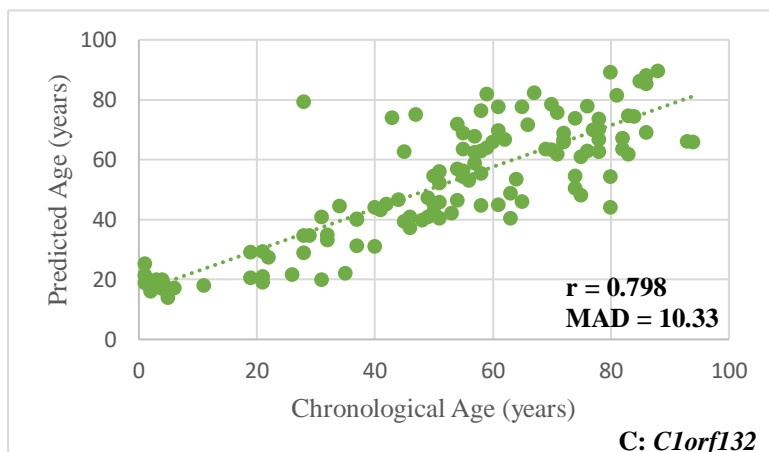
Figure S13: Correlations between DNAm levels and chronological age in 121 blood samples from living and deceased individuals obtained through SNaPshot methodology. A) Positive correlation between methylation levels and chronological age for CpG sites in *ELOVL2* (red), *FHL2* (dark blue), *KLF14* (green) and *TRIM59* (yellow) genes; B) negative correlation between methylation levels and chronological age for *C1orf132* locus (light blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



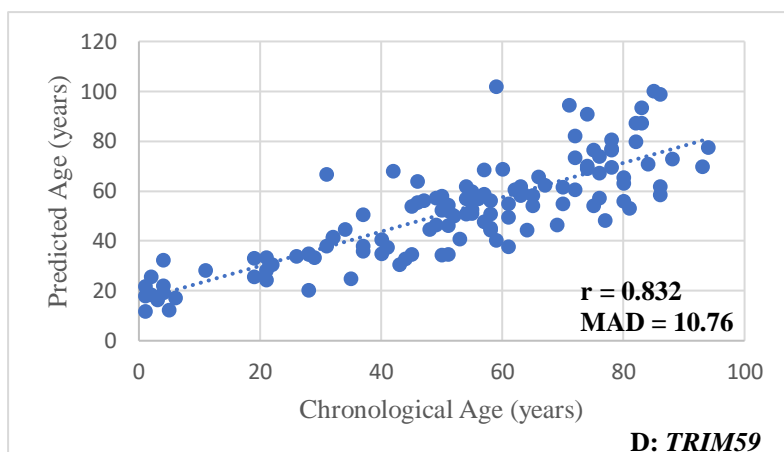
A) $Predicted\ Age = (-3.906) + 181.969 \times DNAm\ level\ ELOVL2.$



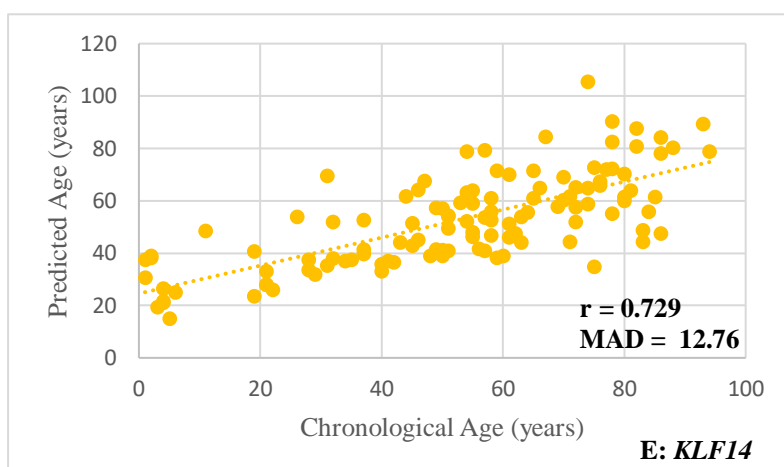
B) $Predicted\ Age = (-3.658) + 179.583 \times DNAm\ level\ FHL2.$



C) $Predicted\ Age = 107.735 - 94.409 \times DNAm\ level\ Clorf132.$



$$D) \text{ Predicted Age} = (-11.700) + 184.634 \times \text{DNAm level TRIM59.}$$



$$E) \text{ Predicted Age} = 14.815 + 554.199 \times \text{DNAm level KLF14.}$$

Figure S14: Plots with predicted age (years) *versus* chronological age (years) of the 121 living and deceased individuals using simple linear regression models developed with the five CpGs in each gene through the multiplex methylation SNaPshot methodology. MAD value and Spearman correlation value (r) are plotted in each chart. A) Predicted age of the 118 individuals based on methylation levels of the CpG from *ELOVL2* gene. For two individuals aged 2 and 5 years old, negative prediction values were obtained and were set at 0; B) Predicted age of the 121 individuals based on methylation levels of the CpG from *FHL2* gene; C) Predicted age of the 117 individuals based on methylation levels of the CpG from *KLF14* gene; D) Predicted age of 119 individuals the based on methylation levels of the CpG from *C1orf132* gene; E) Predicted age of the 121 individuals based on methylation levels of the CpG from *TRIM59* gene.

B. DNA methylation age estimation in tooth samples

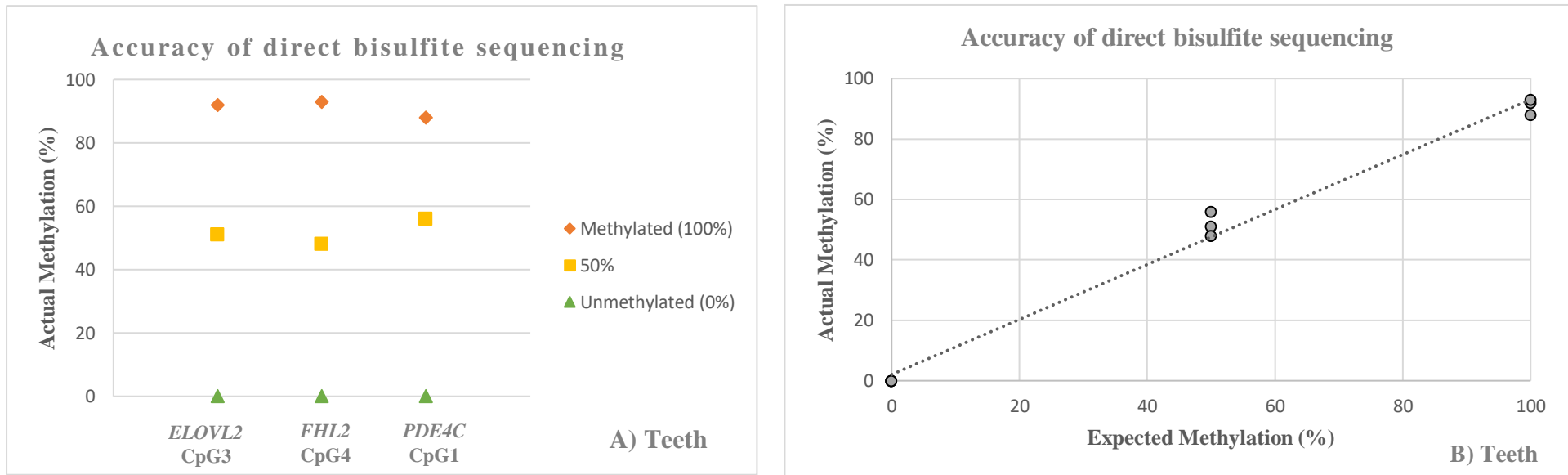


Figure S15: A) Actual methylation (0–100%) obtained for the best-selected CpGs in tooth samples from living and deceased individuals. B) Actual methylation *versus* expected methylation of known quantities of methylated to unmethylated DNA standards.

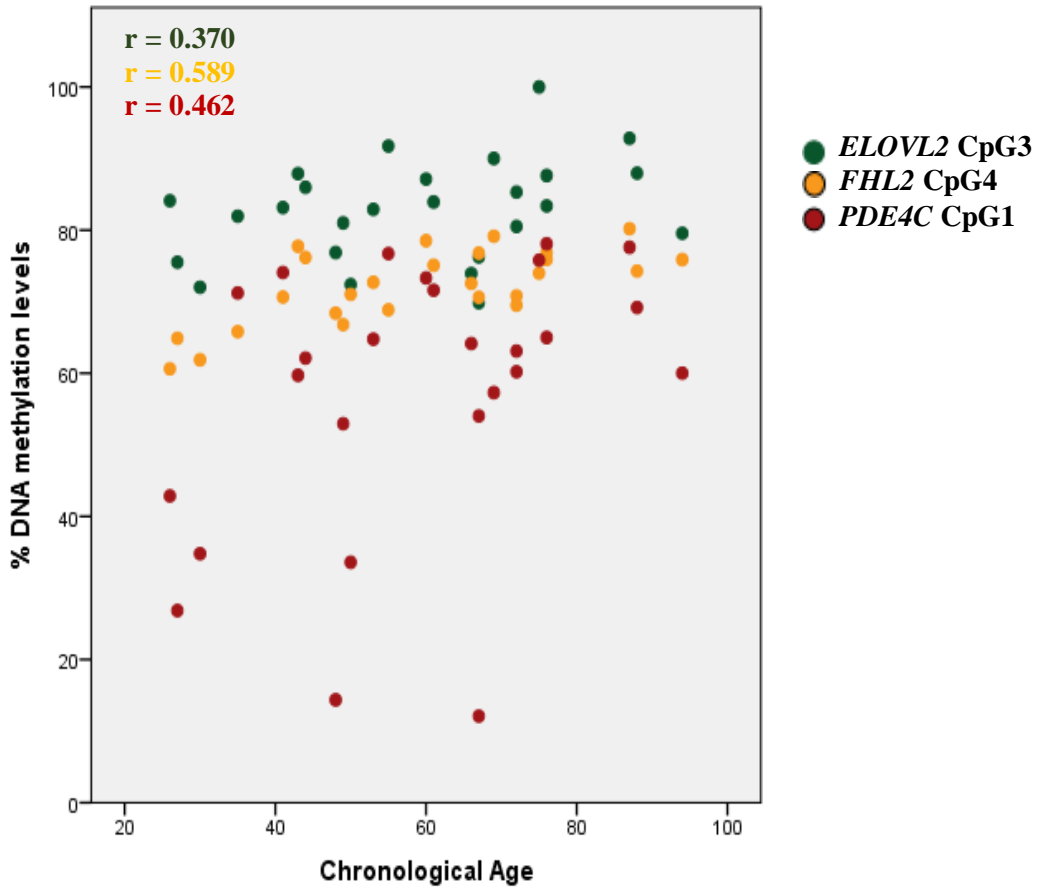
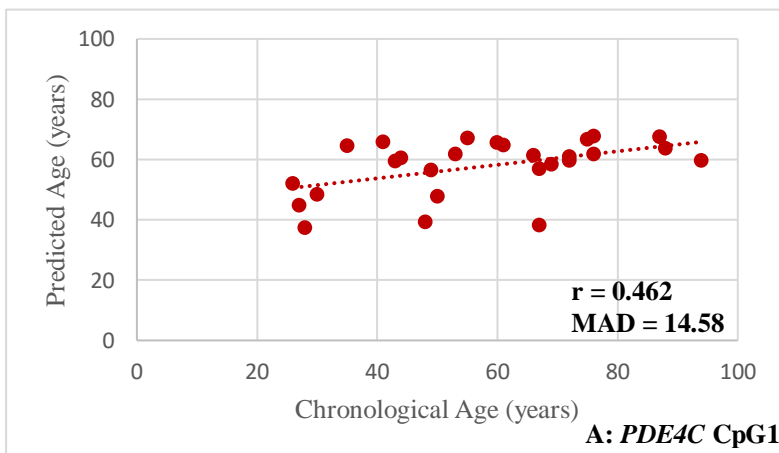
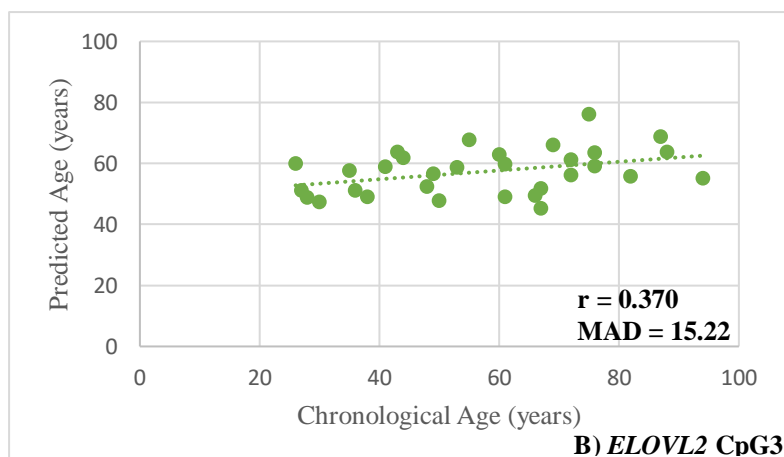


Figure S16: Positive correlation between DNAm levels and chronological age in 31 tooth samples from living and deceased individuals obtained through Sanger sequencing methodology for *ELOVL2* CpG3 (green), *FHL2* CpG4 (yellow) and *PDE4C* CpG1 (dark red) markers. The corresponding Spearman correlation coefficients (r) are depicted inside the plot.



A) $Predicted\ Age = 32.875 + 44.595 \times DNAm\ level\ PDE4C\ CpG1.$



$$B) \text{ Predicted Age} = (-26.616) + 102.809 \times \text{DNAm level } ELOVL2 \text{ CpG3.}$$

Figure S17: Plots with predicted age (years) *versus* chronological age (years) of the 31 living and deceased individuals using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 27 individuals based on the methylation levels of the *PDE4C* CpG1; B) Predicted age based on 31 individuals based on methylation levels of the *ELOVL2* CpG3.

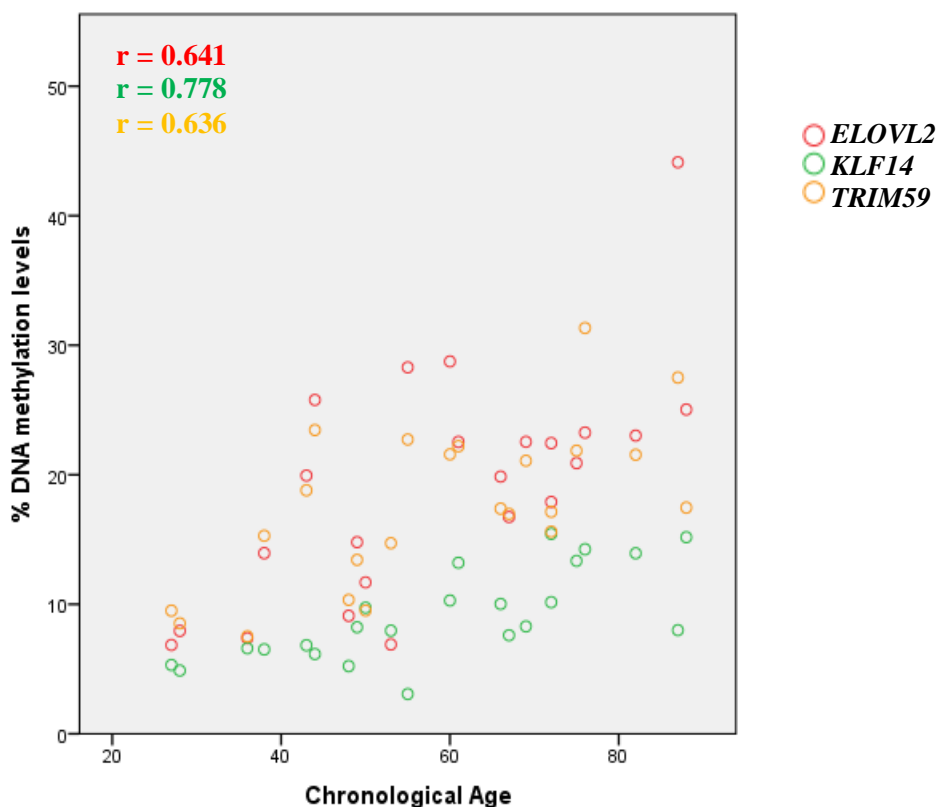
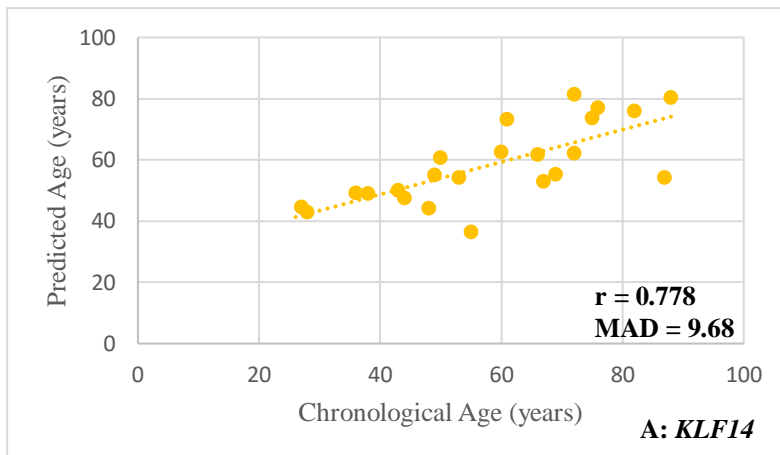
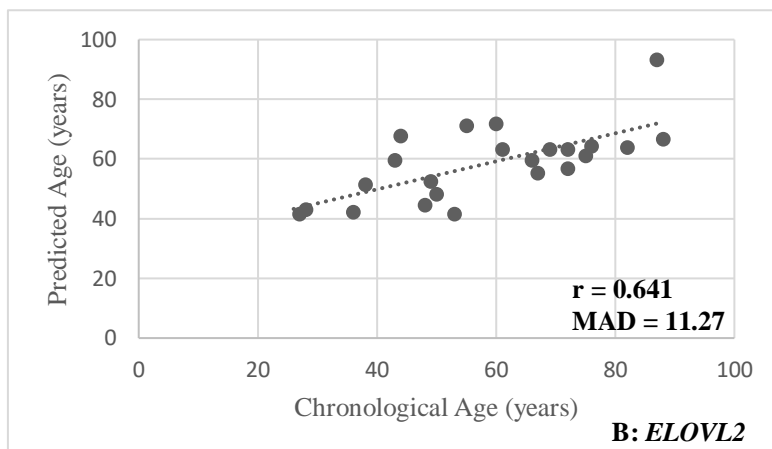


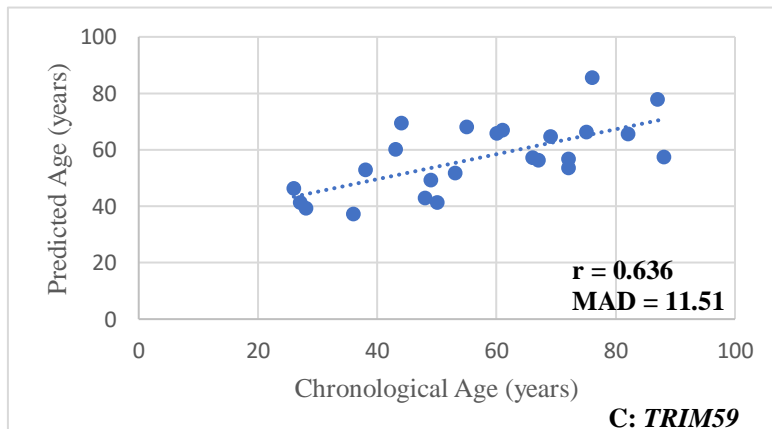
Figure S18: Positive correlation between DNAm levels and chronological age in 24 tooth samples from living and deceased individuals obtained through SNaPshot methodology for CpGs in *ELOVL2* (red), *KLF14* (green) and *TRIM59* (yellow) genes. The corresponding Spearman correlation coefficients (r) are depicted inside the plot.



A) $Predicted\ Age = 25.291 + 363.466 \times DNAm\ level\ KLF14.$



B) $Predicted\ Age = 32.051 + 138.444 \times DNAm\ level\ ELOVL2.$



C) $Predicted\ Age = 21.850 + 203.014 \times DNAm\ level\ TRIM59.$

Figure S19: Plots with predicted age (years) versus chronological age (years) of the 24 living and deceased individuals using simple linear regression models developed with the three CpGs through the multiplex methylation SNaPshot. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 23 individuals based on methylation levels of the CpG from *KLF14* gene; B) Predicted age of the 23 individuals based on methylation levels of the CpG from *ELOVL2* gene; C) Predicted age of the 24 individuals based on methylation levels of the CpG from *TRIM59* gene.

C. DNA methylation age estimation in fresh bone samples

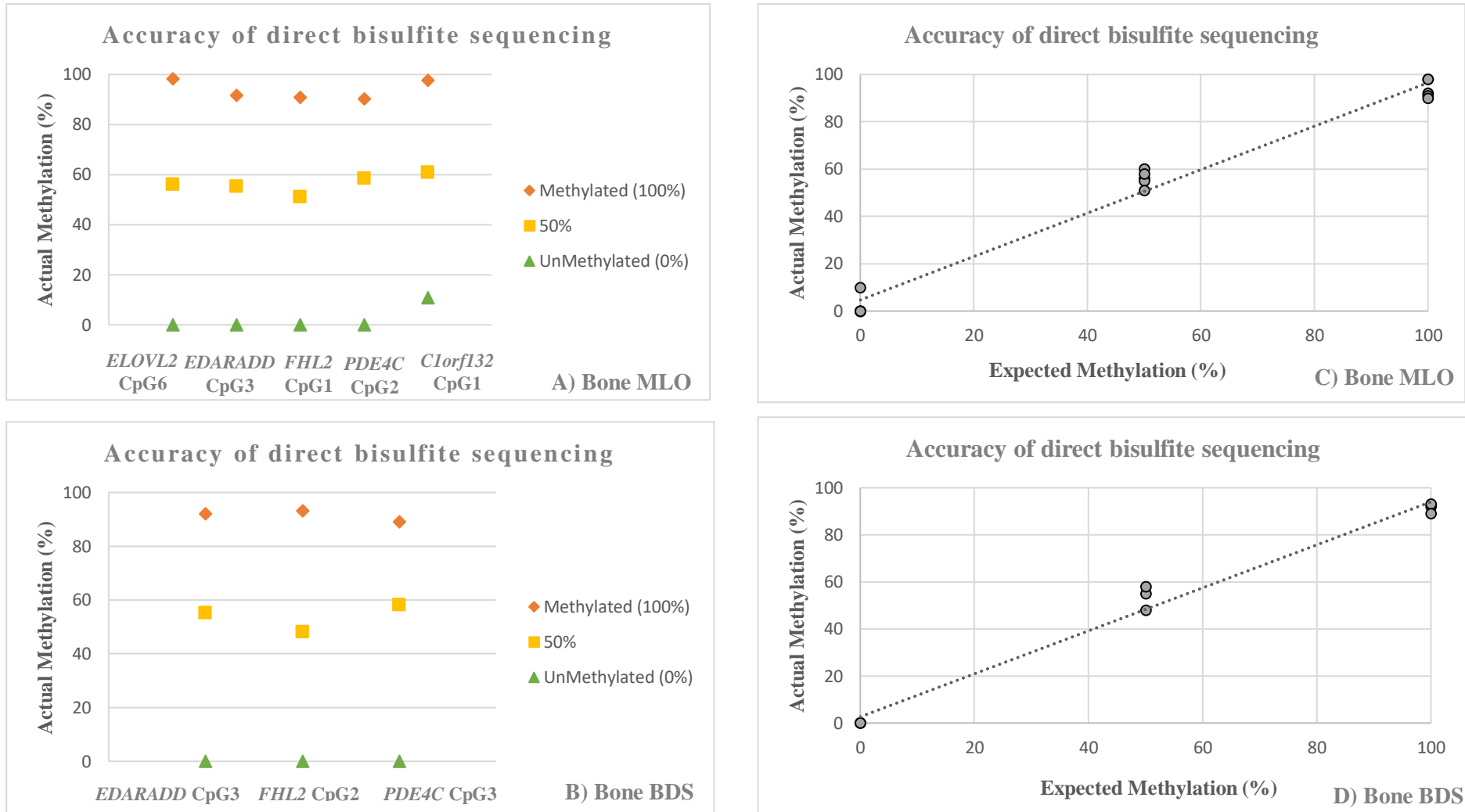


Figure S20: Actual methylation (0–100%) obtained for the best-selected CpGs in bone samples collected during autopsies (A) and in bones collected from BDS (B). Actual methylation *versus* expected methylation of known quantities of methylated to unmethylated DNA standards in bones collected during autopsies (C) and bones collected from BDS (D).

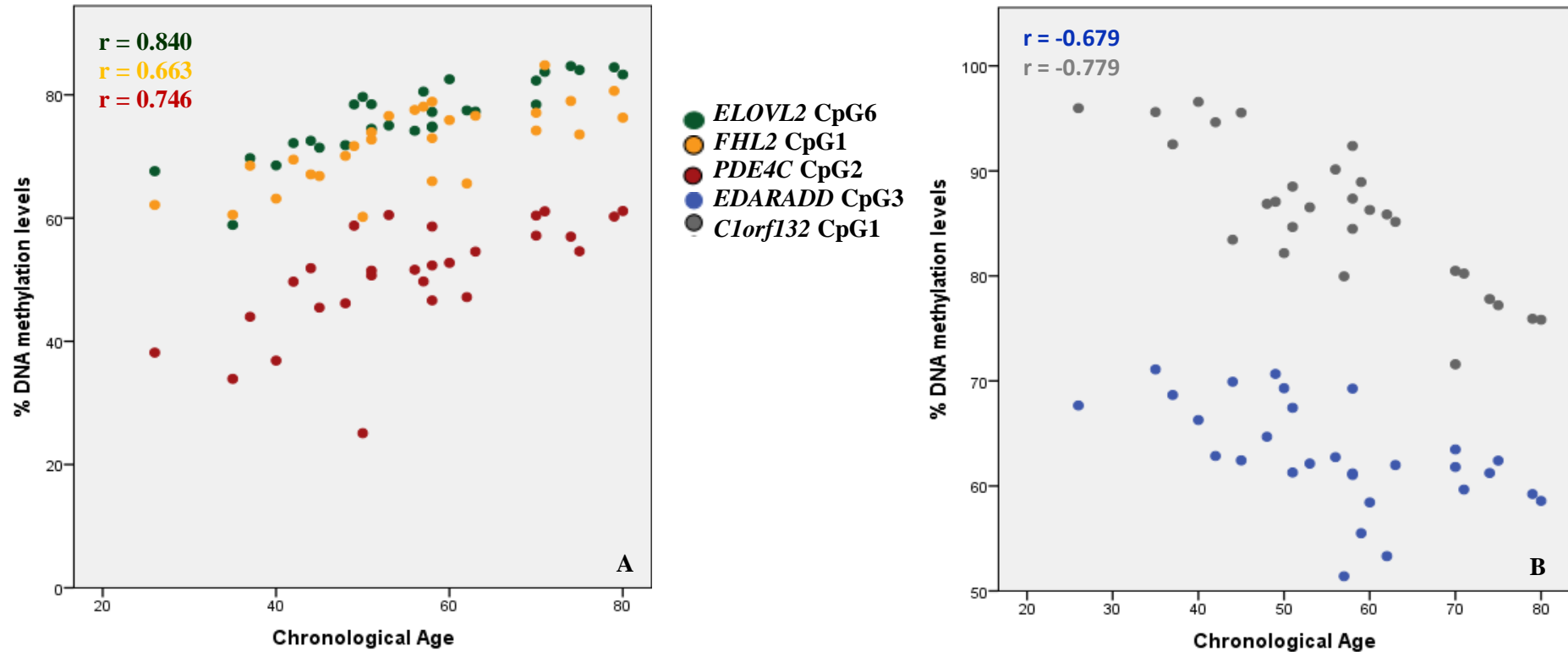
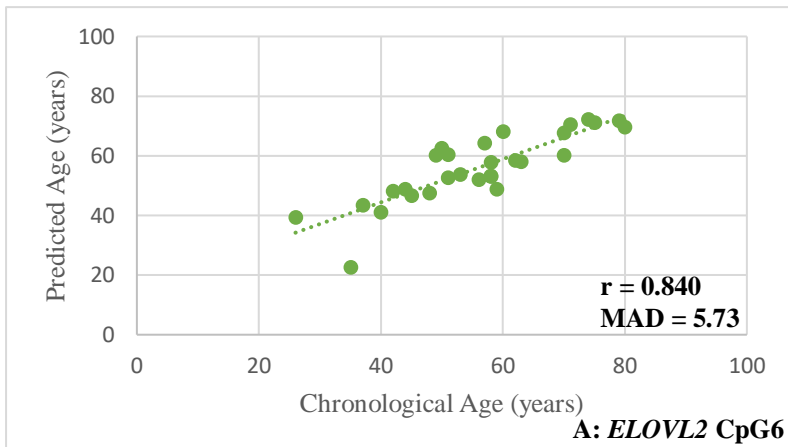
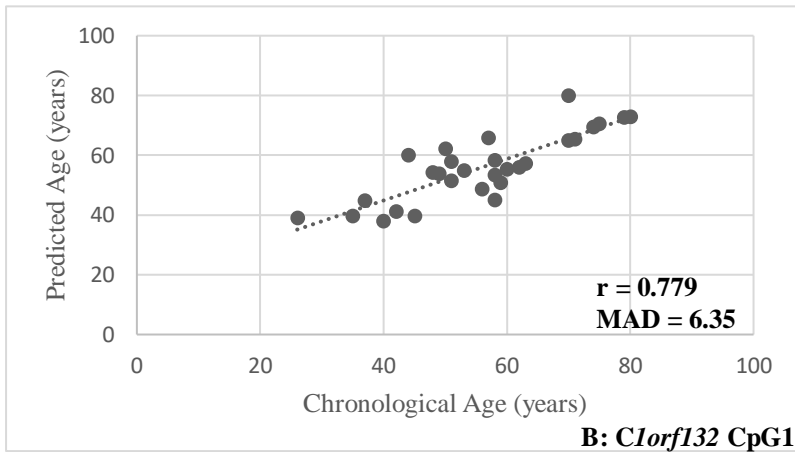


Figure S21: Correlations between DNAm levels and chronological age in 29 fresh bone samples collected during autopsies obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *ELOVL2* CpG6 (green), *FHL2* CpG1 (yellow), *PDE4C* CpG2 (dark red); B) negative correlation between methylation levels and chronological age for *EDARADD* CpG3 (blue) and *Clorf132* CpG1 (gray). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



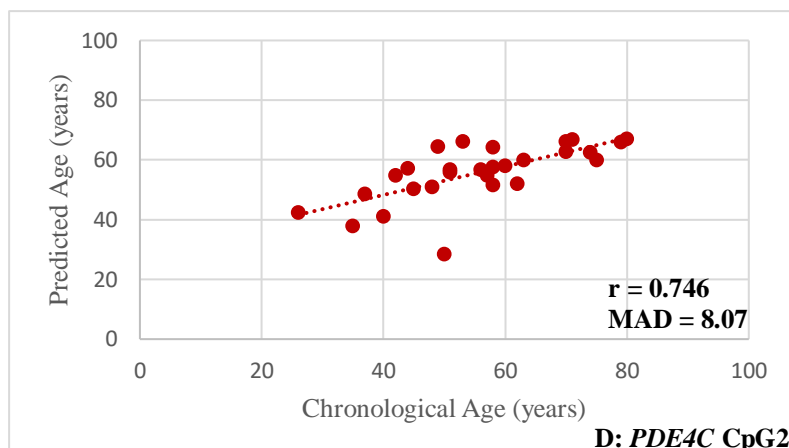
A) $Predicted\ Age = (-91.106) + 192.805 \times DNAm\ level\ ELOVL2\ CpG6.$



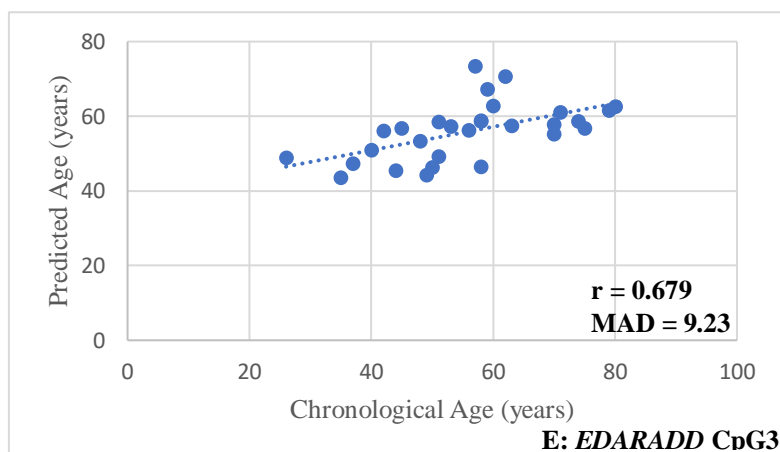
B) $Predicted\ Age = 200.069 - 167.935 \times DNAm\ level\ Clorf132\ CpG1.$



C) $Predicted\ Age = (-48.396) + 145.109 \times DNAm\ level\ FHL2\ CpG1.$



D) $Predicted\ Age = 1.835 + 106.559 \times DNAm\ level\ PDE4C\ CpG2.$



E) $Predicted\ Age = 151.240 - 151.444 \times DNAm\ level\ EDARADD\ CpG3.$

Figure S22: Plots with predicted age (years) *versus* chronological age (years) of the 29 fresh bone samples collected during autopsies using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 29 individuals based on methylation levels of the *ELOVL2* CpG6; B) Predicted age of the 29 individuals based on methylation levels of the *C1orf132* CpG1; C) Predicted age of the 29 individuals based on methylation levels of the *FHL2* CpG1; D) Predicted age of the 28 individuals based on methylation levels of the *PDE4C* CpG2; E) Predicted age of the 29 individuals based on methylation levels of the *EDARADD* CpG3.

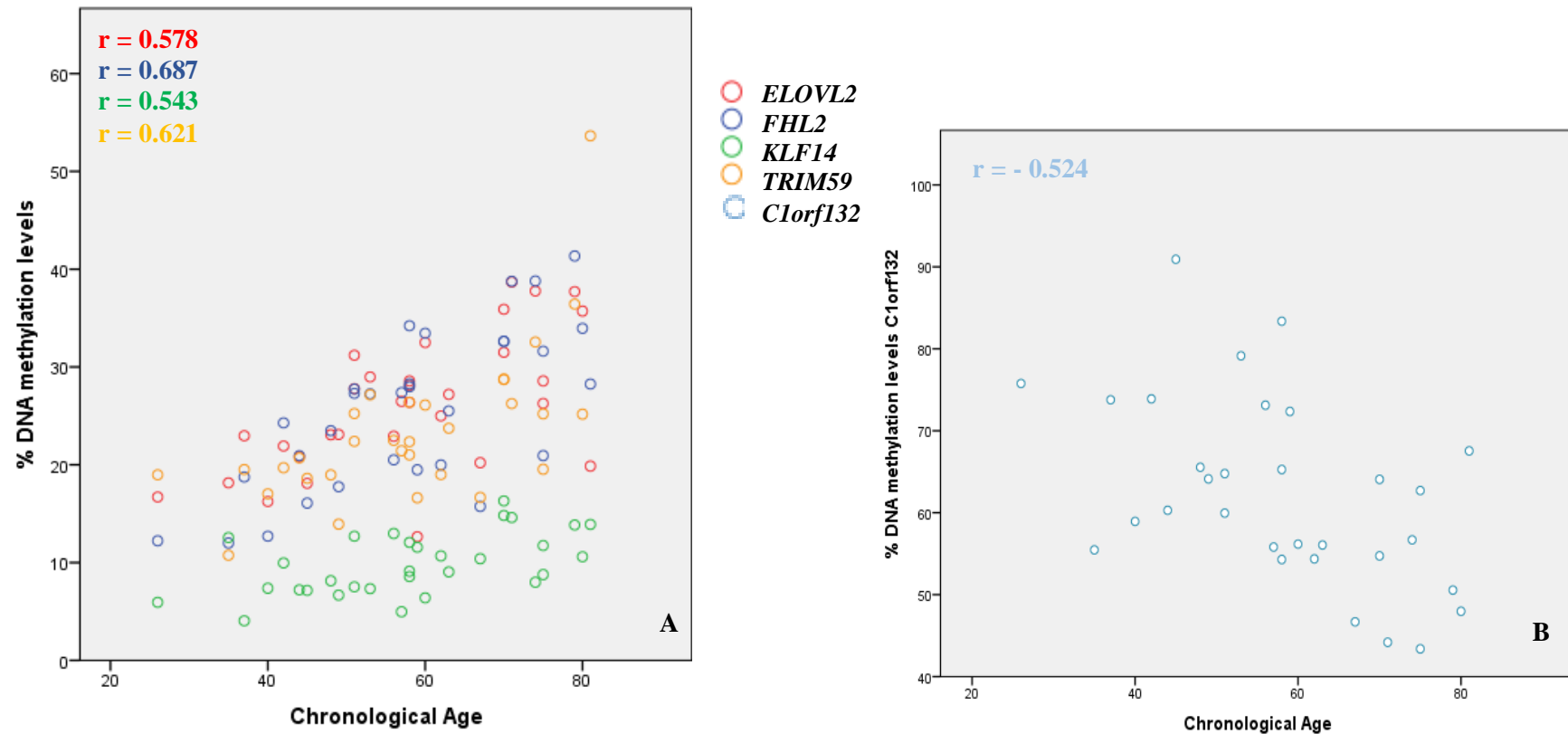
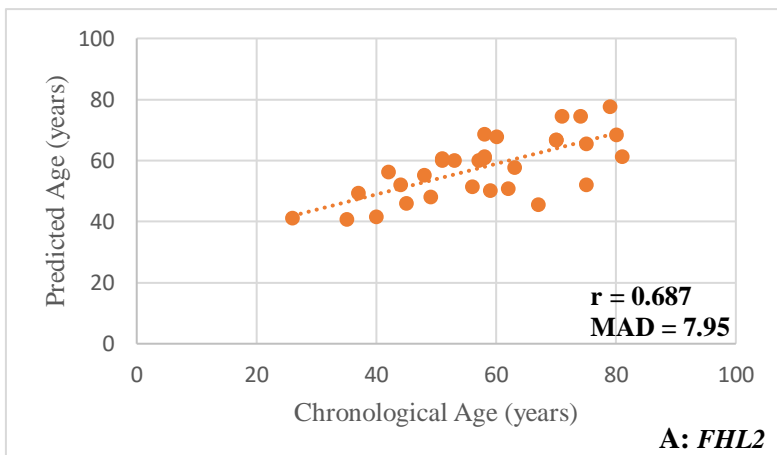
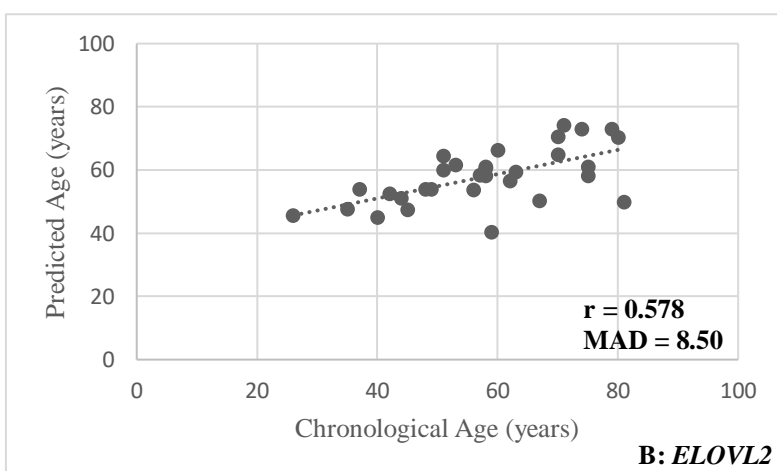


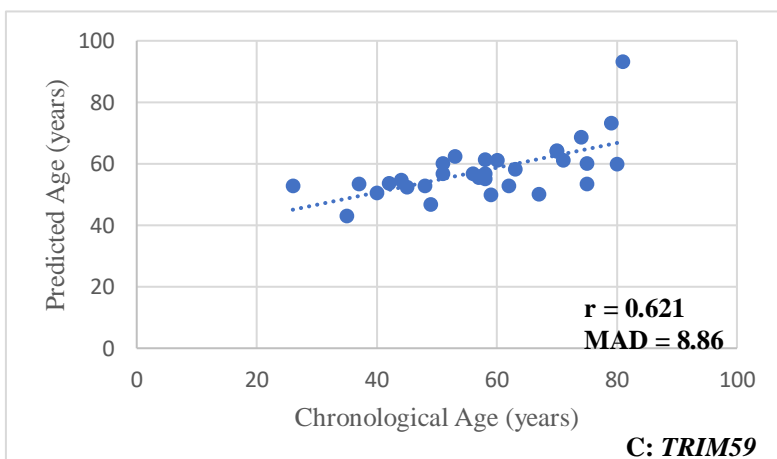
Figure S23: Correlations between DNAm levels and chronological age in the 31 fresh bone samples collected during autopsies obtained through SNaPshot methodology. A) Positive correlation between methylation levels and chronological age for CpG sites in *ELOVL2* (red), *FHL2* (blue), *KLF14* (green) and *TRIM59* (yellow) genes; B) negative correlation between methylation levels and chronological age for CpG in *C1orf132* locus (light blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



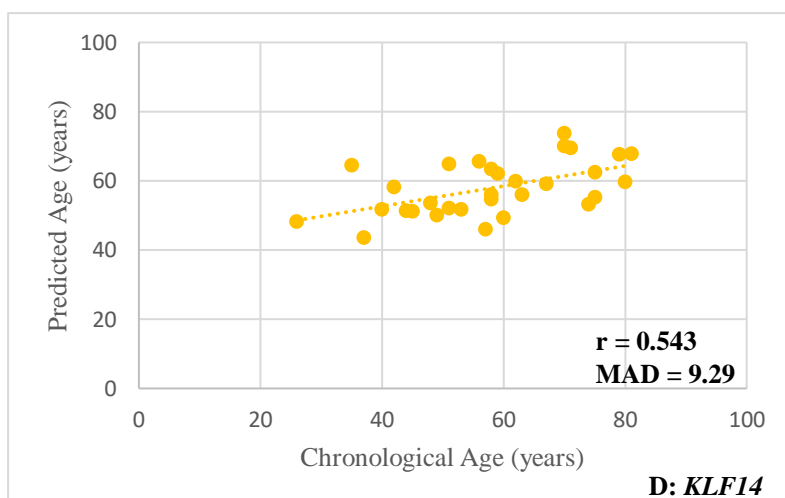
A) $\text{Predicted Age} = 25.749 + 125.744 \times \text{DNAm level FHL2}$.



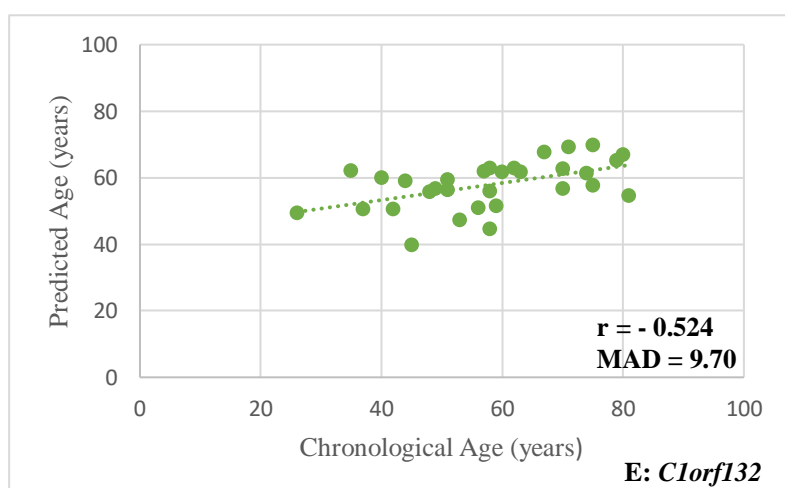
B) $\text{Predicted Age} = 23.916 + 129.770 \times \text{DNAm level ELOVL2}$.



C) $\text{Predicted Age} = 30.532 + 116.880 \times \text{DNAm level TRIM59}$.



D) $Predicted\ Age = 33.732 + 245.312 \times DNAm\ level\ KLF14.$



E) $Predicted\ Age = 97.182 - 63.078 \times DNAm\ level\ Clorf132.$

Figure S24: Plots with predicted age (years) *versus* chronological age (years) of the 31 bone samples from deceased individuals collected during autopsies using simple linear regression models developed with the five CpGs in each gene through the multiplex methylation SNaPshot. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 31 individuals based on methylation levels of the CpG from *ELOVL2* gene; B) Predicted age of the 31 individuals based on CpG from *FHL2* gene; C) Predicted age of the 31 individuals based on methylation levels of the CpG from *KLF14* gene; D) Predicted age of the 31 individuals based on methylation levels of the CpG from *Clorf132* gene; E) Predicted age of the 31 individuals based on methylation levels of the CpG from *TRIM59* gene.

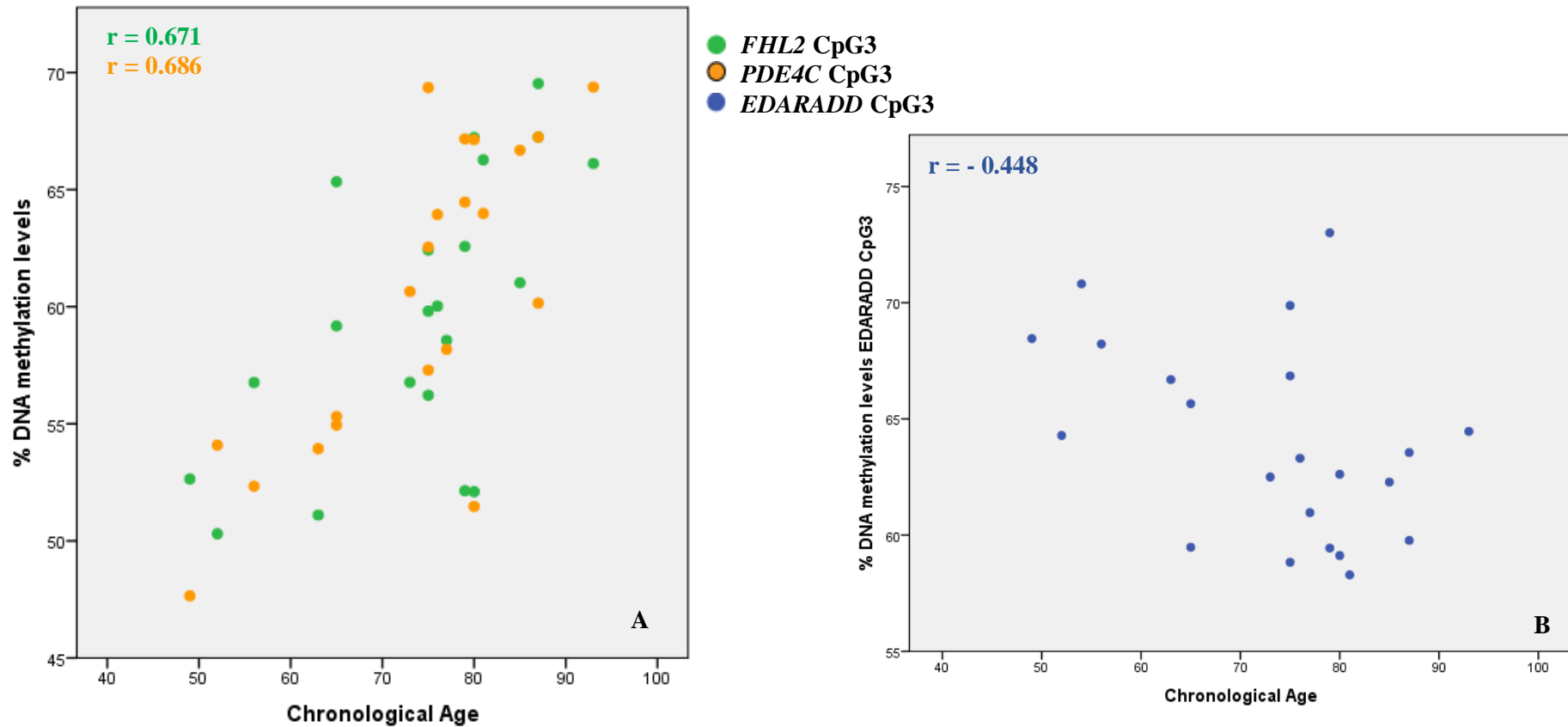
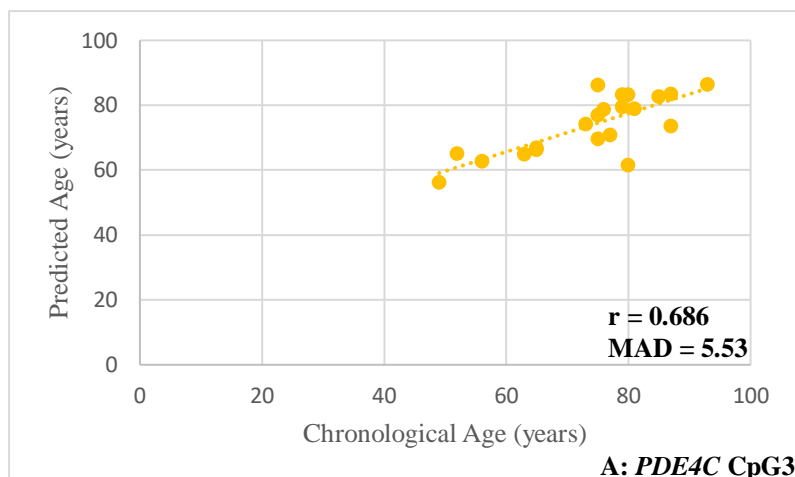
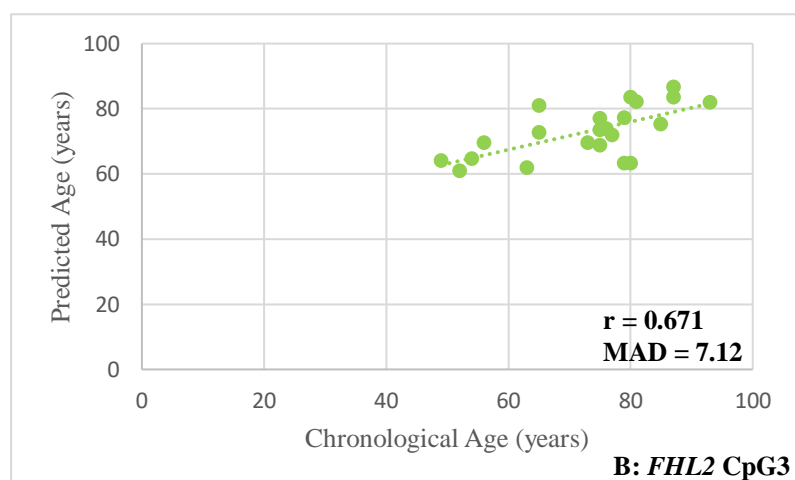


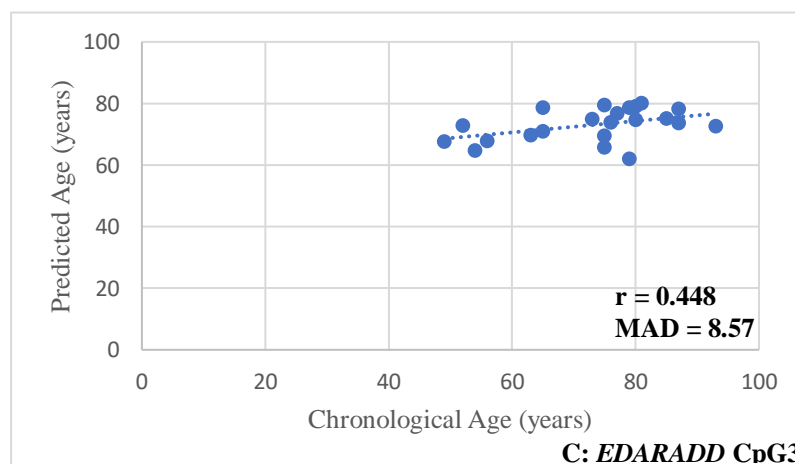
Figure S25: Correlations between DNAm levels and chronological age in 22 bone samples collected from BDS obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *FHL2* CpG3 (green) and *PDE4C* CpG3 (yellow); B) negative correlation between methylation levels and chronological age for *EDARADD* CpG3 (blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



$$\text{A) Predicted Age} = (-9.685) + 138.451 \times \text{DNAm level } PDE4C \text{ CpG3.}$$



$$\text{B) Predicted Age} = (-6.427) + 133.754 \times \text{DNAm level } FHL2 \text{ CpG3}$$



$$\text{C) Predicted Age} = 151.880 - 123.208 \times \text{DNAm level } EDARADD \text{ CpG3.}$$

Figure S26: Plots with predicted age (years) *versus* chronological age (years) of 22 bone samples collected from BDS using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 21 individuals based on methylation levels of the *PDE4C* CpG3; B) Predicted age of the 22 individuals based on methylation levels of the *FHL2* CpG3; C) Predicted age of the 22 individuals based on methylation levels of the *EDARADD* CpG3.

D. DNA methylation age estimation in buccal swabs

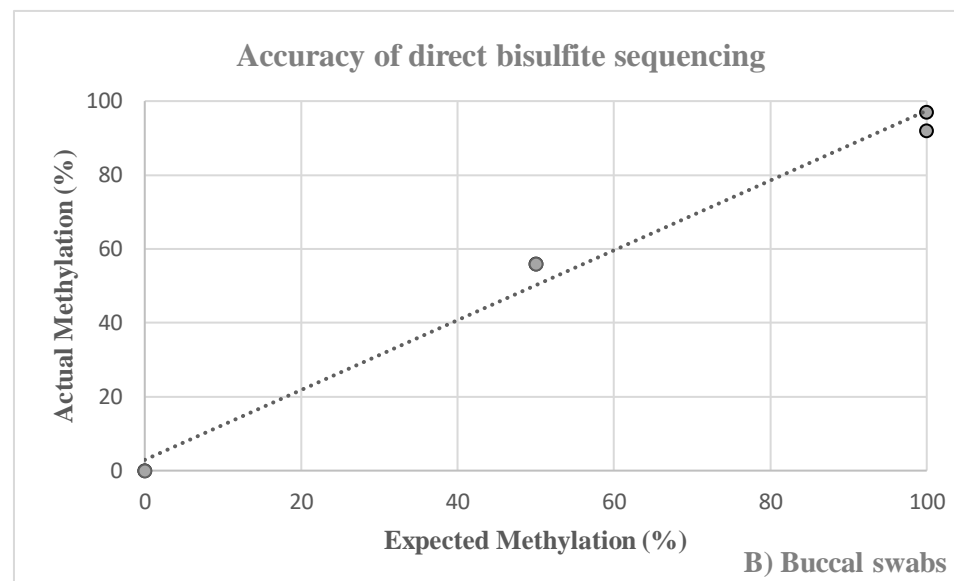
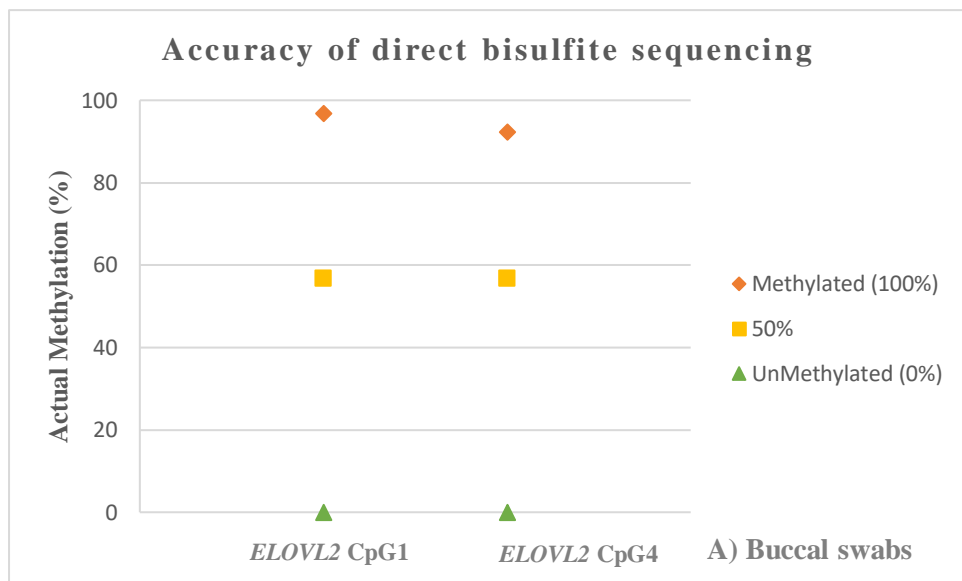


Figure S27: A) Actual methylation (0–100%) obtained for the best-selected CpGs in buccal swabs from living individuals. B) Actual methylation *versus* expected methylation of known quantities of methylated to unmethylated DNA standards.

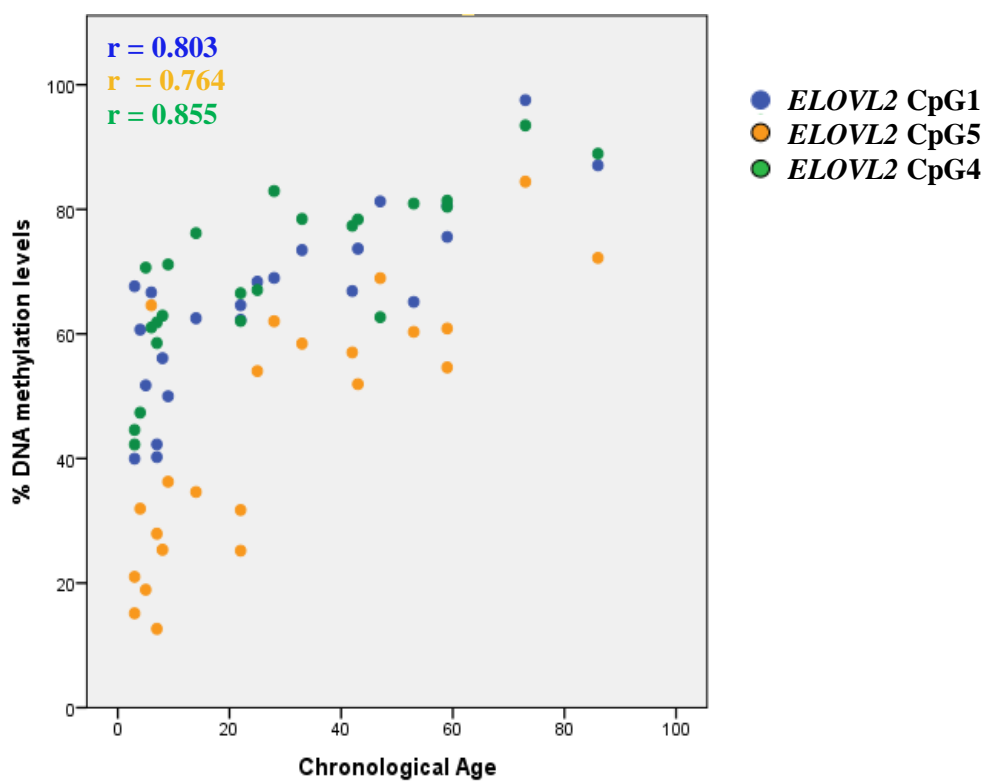
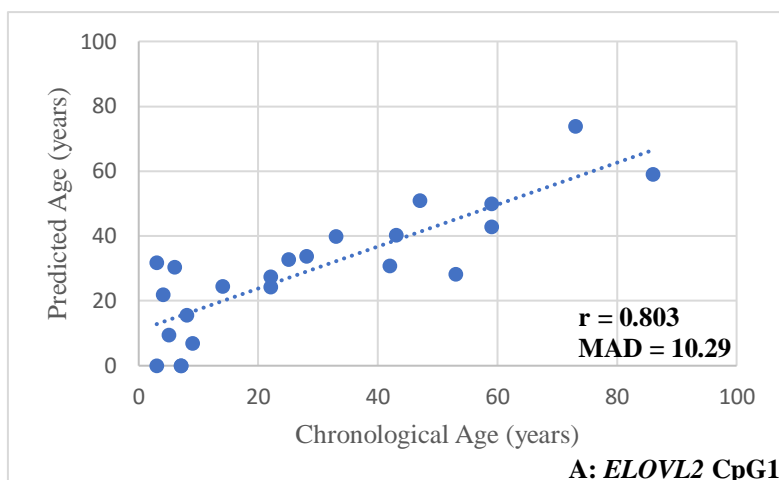
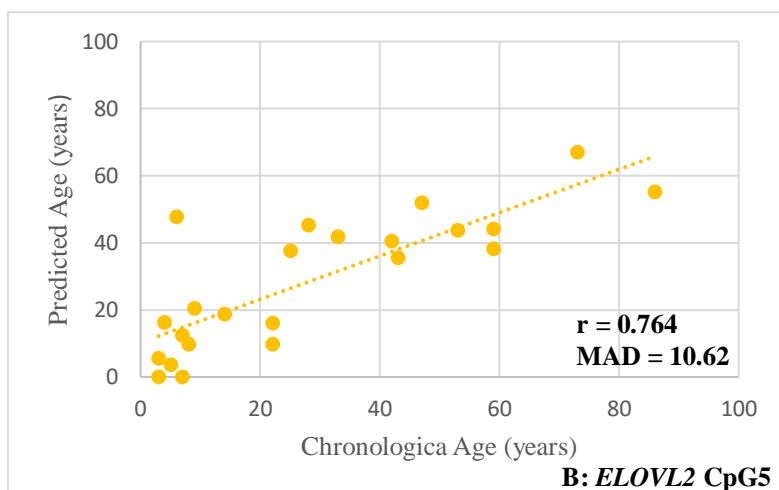


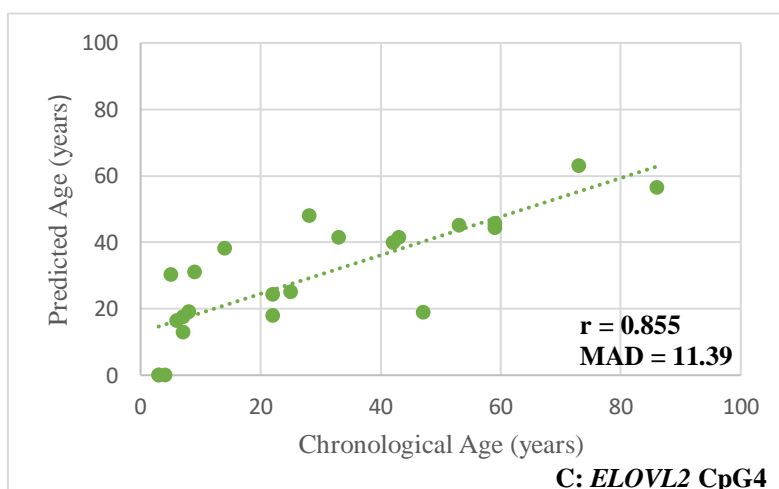
Figure S28: Positive correlation between DNAm levels and chronological age in 23 buccal swabs from living individuals obtained through Sanger sequencing methodology for *ELOVL2* CpG1 (blue), *ELOVL2* CpG5 (yellow) and *ELOVL2* CpG4 (green). The corresponding Spearman correlation coefficients (r) are depicted inside the plot.



A) $Predicted\ Age = (-63.327) + 140.654 \times DNAm\ level\ ELOVL2\ CpG1.$



B) $Predicted\ Age = (-14.728) + 96.741 \times DNAm\ level\ ELOVL2\ CpG5.$



C) $Predicted\ Age = (-71.024) + 143.469 \times DNAm\ level\ ELOVL2\ CpG4.$

Figure S29: Plots with predicted age (years) *versus* chronological age (years) of the 23 buccal swabs from living individuals using simple linear regression models developed with the best CpGs in *ELOVL2* gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age based on methylation levels of *ELOVL2* CpG1; B) Predicted age based on methylation levels of *ELOVL2* CpG5; C) Predicted age based on methylation levels of *ELOVL2* CpG4.

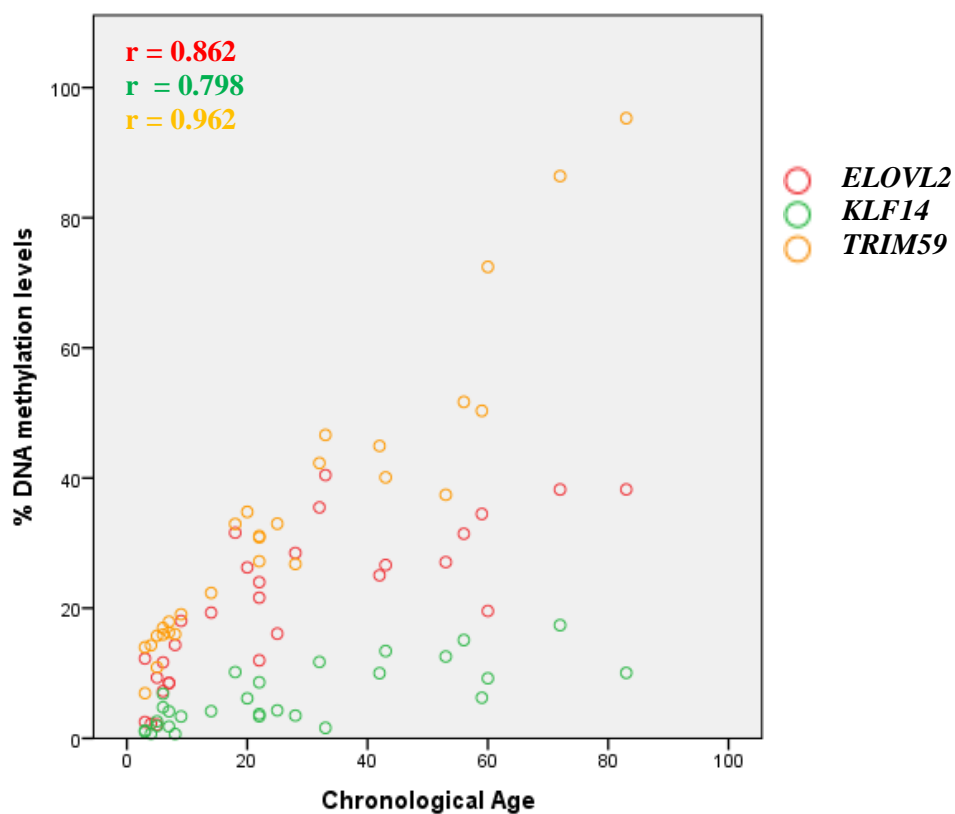
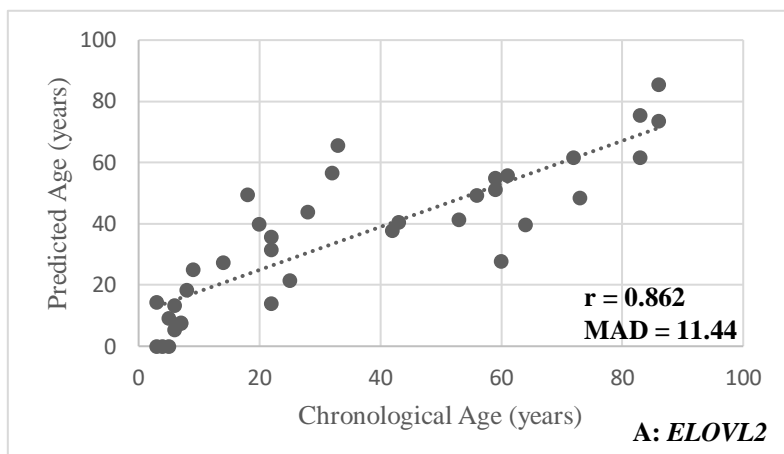
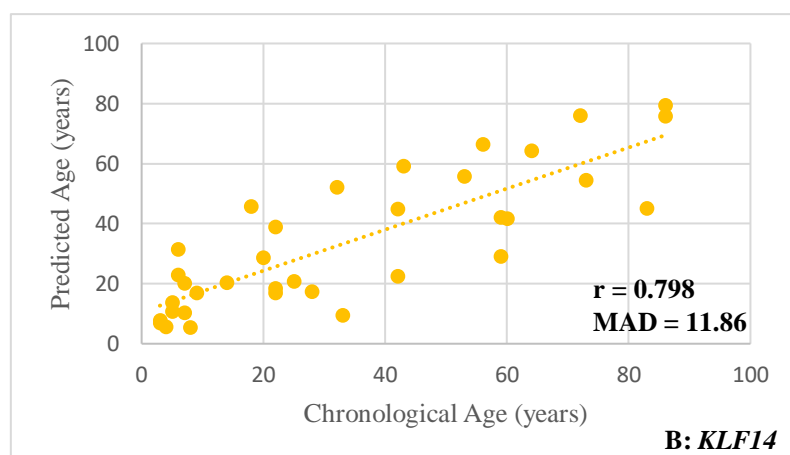


Figure S30: Positive correlation between DNAm levels and chronological age in 39 buccal swabs from living individuals obtained through SNaPshot methodology for CpG sites in *ELOVL2* (red), *KLF14* (green) and *TRIM59* (yellow) genes. The corresponding Spearman correlation coefficients (r) are depicted inside the plot.



$$A) \text{ Predicted Age} = (-7.867) + 181.842 \times \text{DNAm level}$$



$$B) \text{ Predicted Age} = 2.725 + 421.256 \times \text{DNAm level}$$

Figure S31: Plots with predicted age (years) *versus* chronological age (years) of the 39 buccal swabs using simple linear regression models developed with the three CpGs in each gene through the multiplex methylation SNaPshot. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 36 individuals based on methylation levels of the CpG from *ELOVL2* gene; B) Predicted age of the 35 the individuals based on methylation levels of the CpG from *KLF14* gene.

E. DNA methylation age estimation through multi-tissues

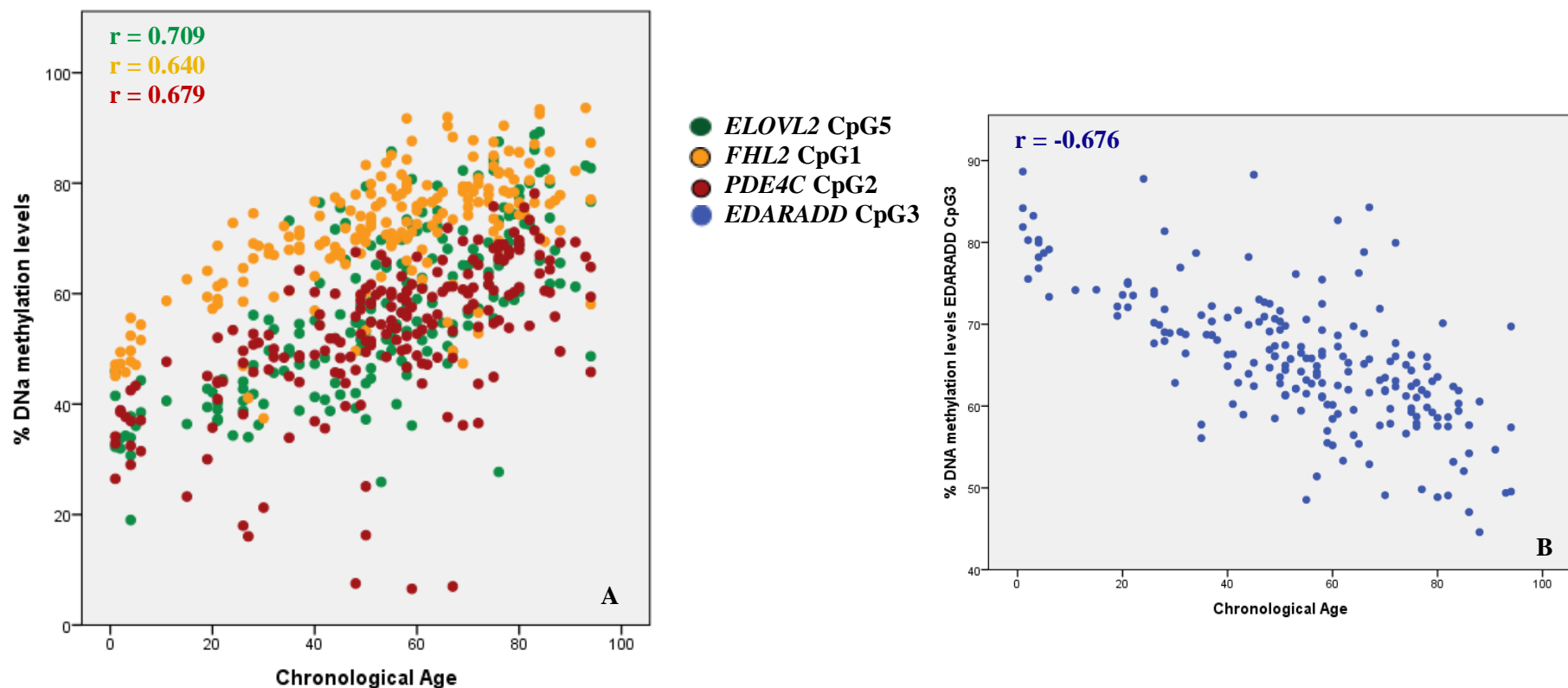
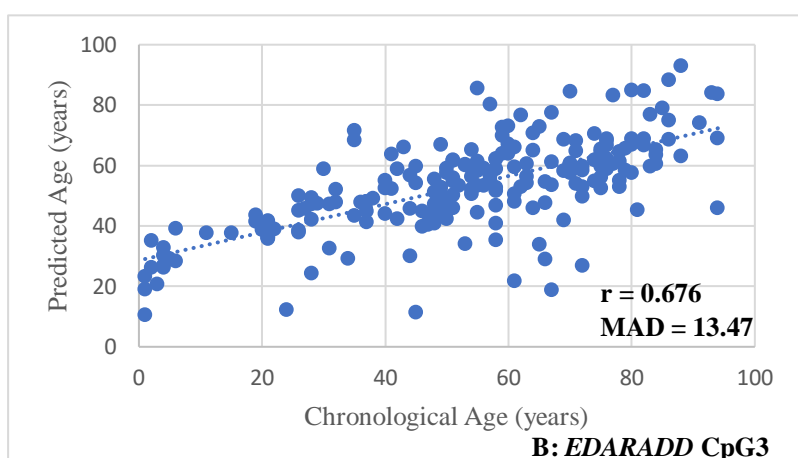


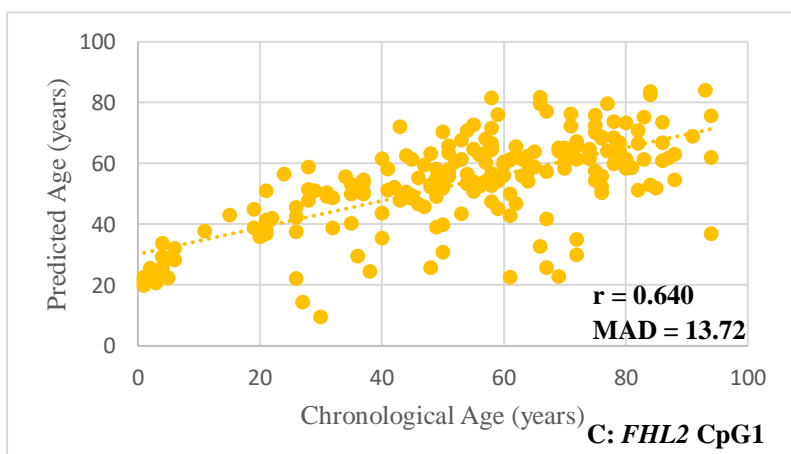
Figure S32: Correlations between DNAm levels and chronological age in 204 samples included in group 1 (including blood samples from living and deceased individuals, tooth from living and deceased and bone samples collected from autopsies) obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *ELOVL2* CpG5 (green), *PDE4C* CpG2 (dark red), *FHL2* CpG1 (yellow) markers; B) negative correlation between methylation levels and chronological age for *EDARADD* CpG3 (blue) markers. The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



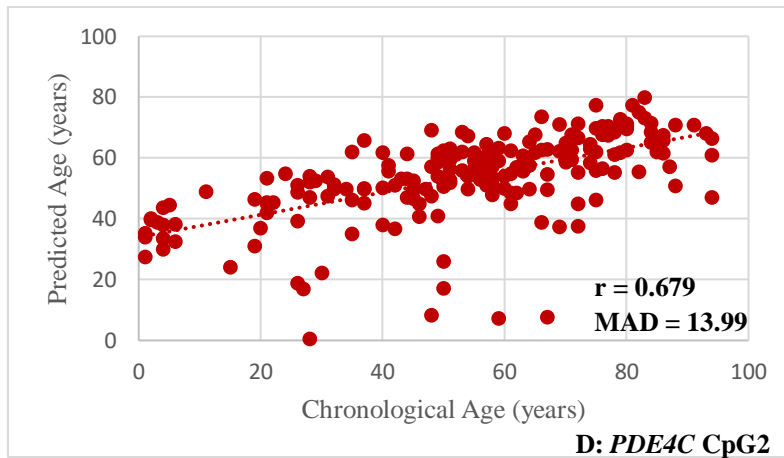
A) $Predicted\ Age = (-7.704) + 107.746 \times DNAm\ level\ ELOVL2\ CpG5.$



B) $Predicted\ Age = 176.416 - 186.873 \times DNAm\ level\ EDARADD\ CpG3.$



C) $Predicted\ Age = (-40.019) + 132.373 \times DNAm\ level\ FHL2\ CpG1.$



$$D) \text{ Predicted Age} = 0.468 + 101.410 \times \text{DNAm level PDE4C CpG2.}$$

Figure S33: Plots with predicted age (years) *versus* chronological age (years) of the 204 samples, including blood samples from living and deceased individuals, tooth from living and deceased and bone samples collected from autopsies, using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 201 individuals based on methylation levels of the *ELOVL2* CpG5; B) Predicted age of the 202 individuals based on methylation levels of *EDARADD* CpG3; C) Predicted age of the 203 individuals based on methylation levels of the *FHL2* CpG1; D) Predicted age of the 197 individuals based on methylation levels of the *PDE4C* CpG2.

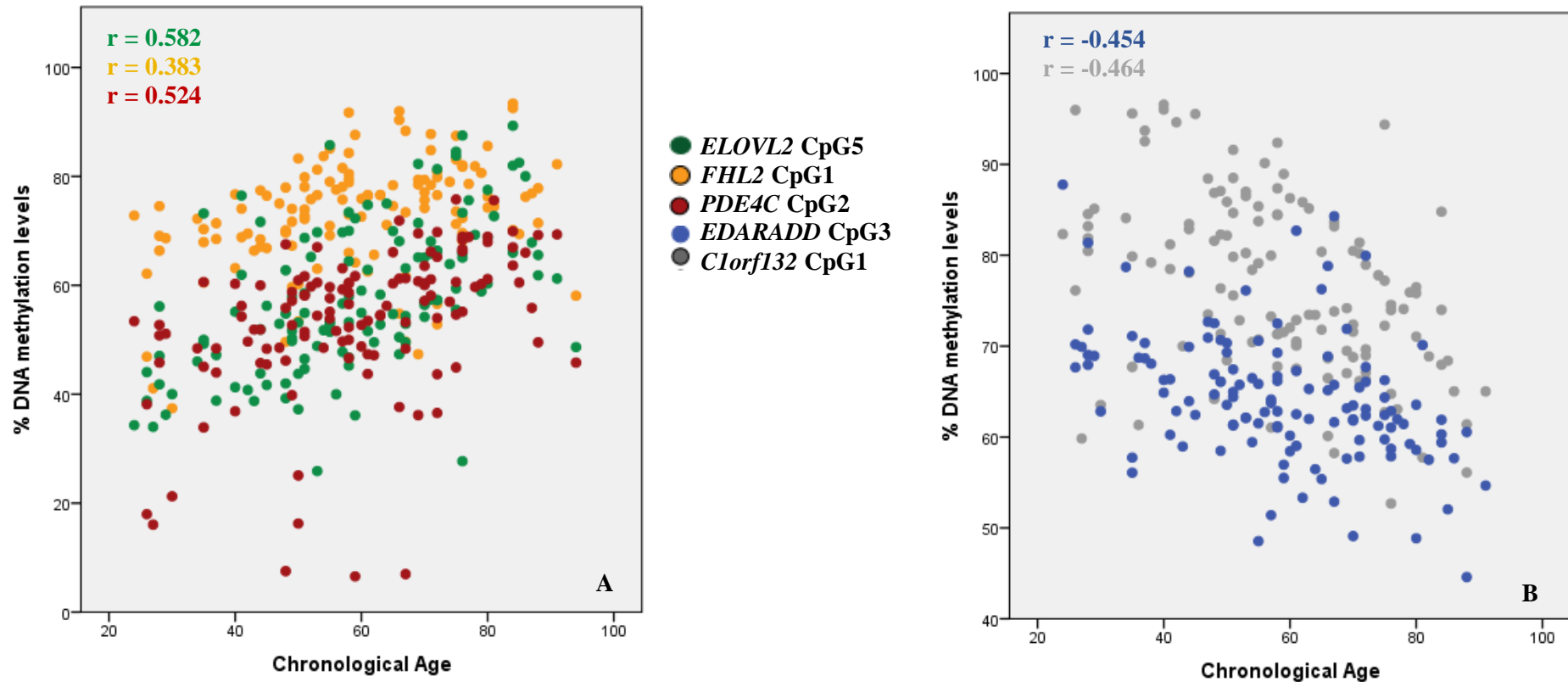
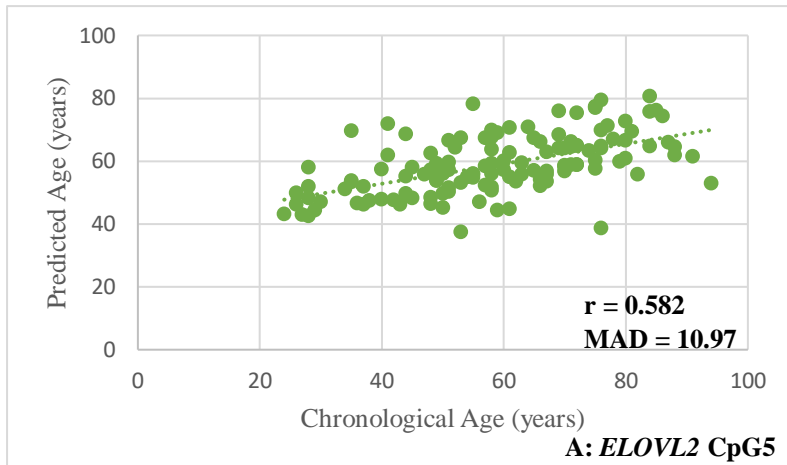
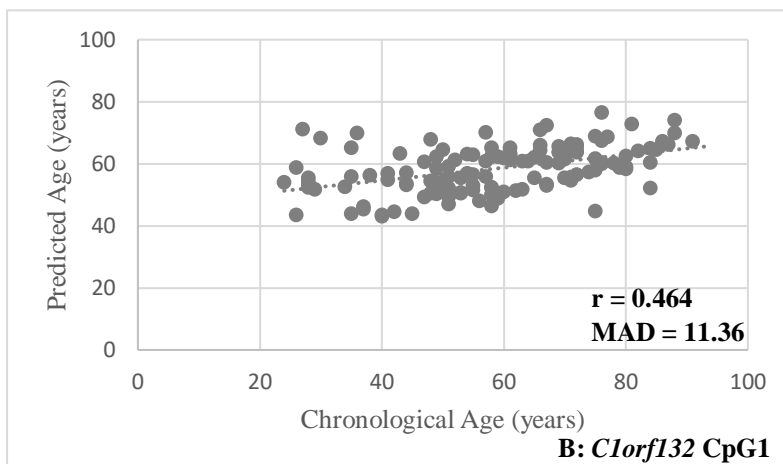


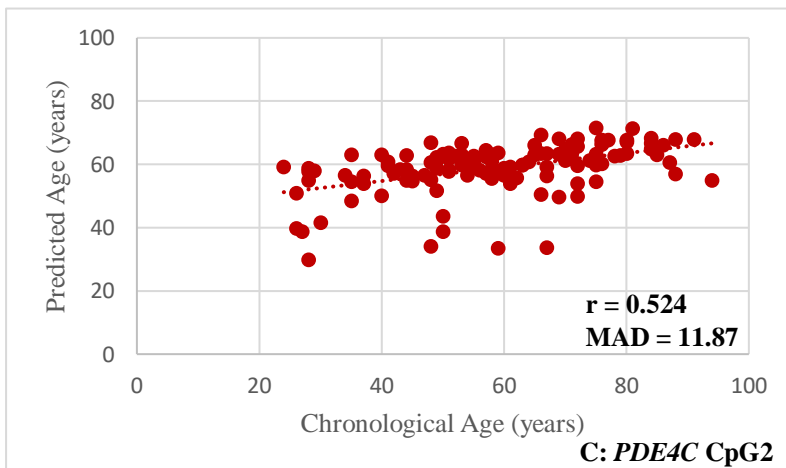
Figure S34: Correlations between DNAm levels and chronological age in 133 samples included in group 2 (including blood from deceased individuals, tooth from living and deceased individuals and bone samples collected from autopsies) obtained through Sanger sequencing methodology. A) Positive correlation between methylation levels and chronological age for *ELOVL2* CpG5 (green), *PDE4C* CpG2 (dark red), *FHL2* CpG1 (yellow) markers; B) negative correlation between methylation levels and chronological age for *Clorf132* CpG1 (gray) and *EDARADD* CpG3 (blue) markers. The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



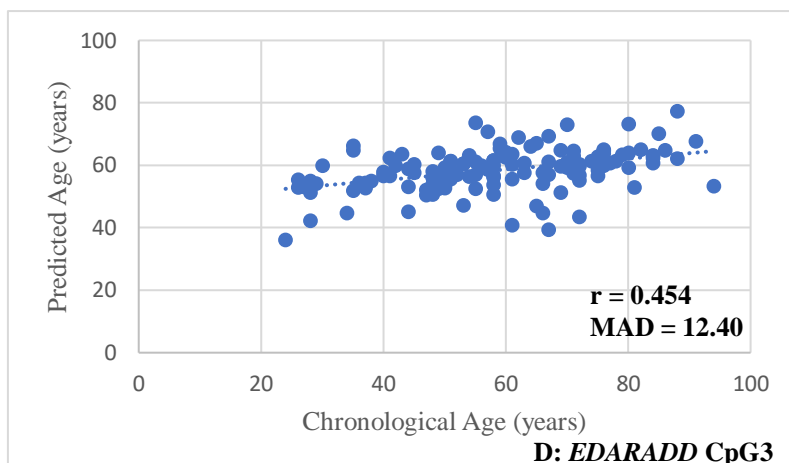
A) $Predicted\ Age = 19.663 + 68.308 \times DNAm\ level\ ELOVL2\ CpG5.$



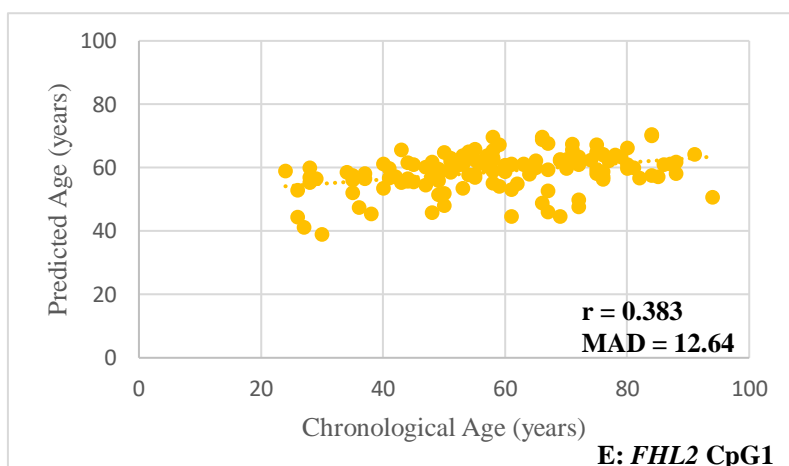
B) $Predicted\ Age = 116.791 - 76.247 \times DNAm\ level\ C1orf132\ CpG1.$



C) $Predicted\ Age = 29.828 + 54.832 \times DNAm\ level\ PDE4C\ CpG2.$



D) $Predicted\ Age = 119.899 - 95.515 \times DNAm\ level\ EDARADD\ CpG3.$



E) $Predicted\ Age = 17.967 + 56.274 \times DNAm\ level\ FHL2\ CpG1.$

Figure S35: Plots with predicted age (years) *versus* chronological age (years) of the 133 samples, including blood from deceased individuals, tooth from living and deceased individuals and bone samples collected from autopsies, using simple linear regression models developed with the best CpG site in each gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 130 individuals based on methylation levels of the *ELOVL2* CpG5; B) Predicted age of the 131 individuals based on methylation levels of *EDARADD* CpG3; C) Predicted age of the 132 individuals based on methylation levels of the *FHL2* CpG1; D) Predicted age of the 126 individuals based on methylation levels of the *PDE4C* CpG2.

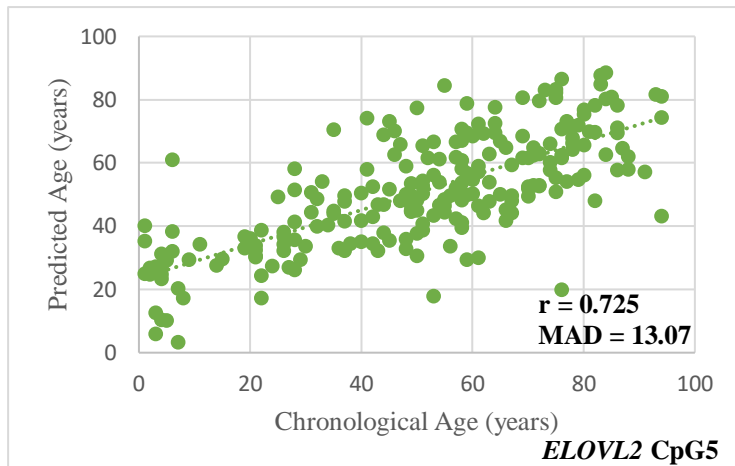


Figure S36: Plot with predicted age (years) *versus* chronological age (years) of the 224 samples, including blood samples from living and deceased individuals, tooth from living and deceased individuals, buccal swabs and bone collected during autopsies, using simple linear regression models developed with the best CpG site in *ELOVL2* gene through the Sanger sequencing methodology. MAD value and Spearman correlation coefficient (r) are plotted in the chart.

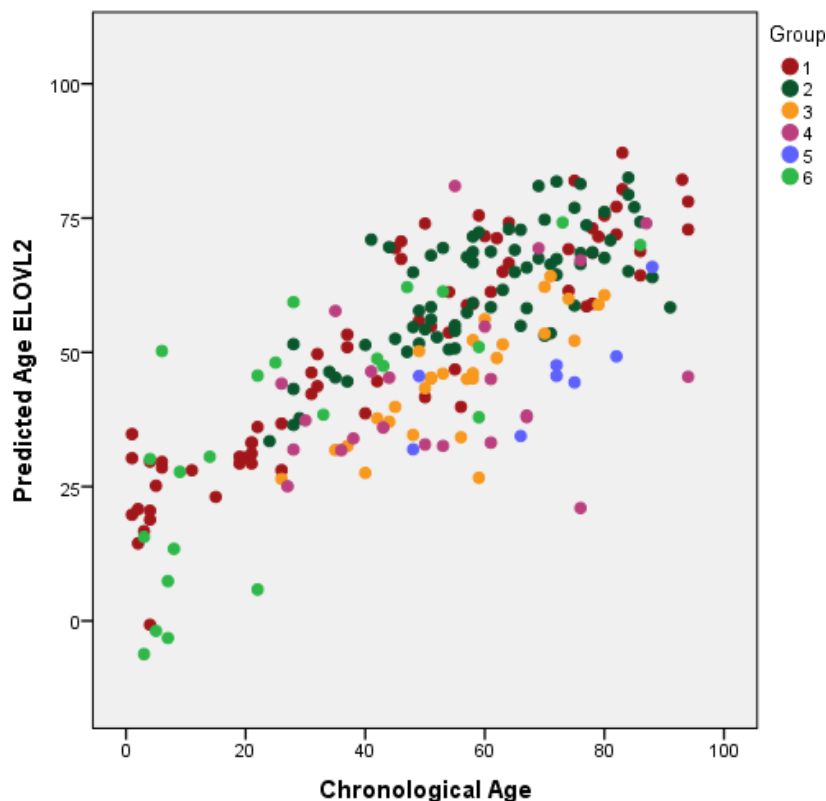


Figure S37: Predicted *versus* chronological ages using the one locus multi-tissue model developed for *ELOVL2* gene included blood samples from living (1), blood samples from deceased individuals (2), bone samples collected from autopsies (3), tooth samples from living (4), tooth samples from deceased individuals (5) and buccal swabs (6).

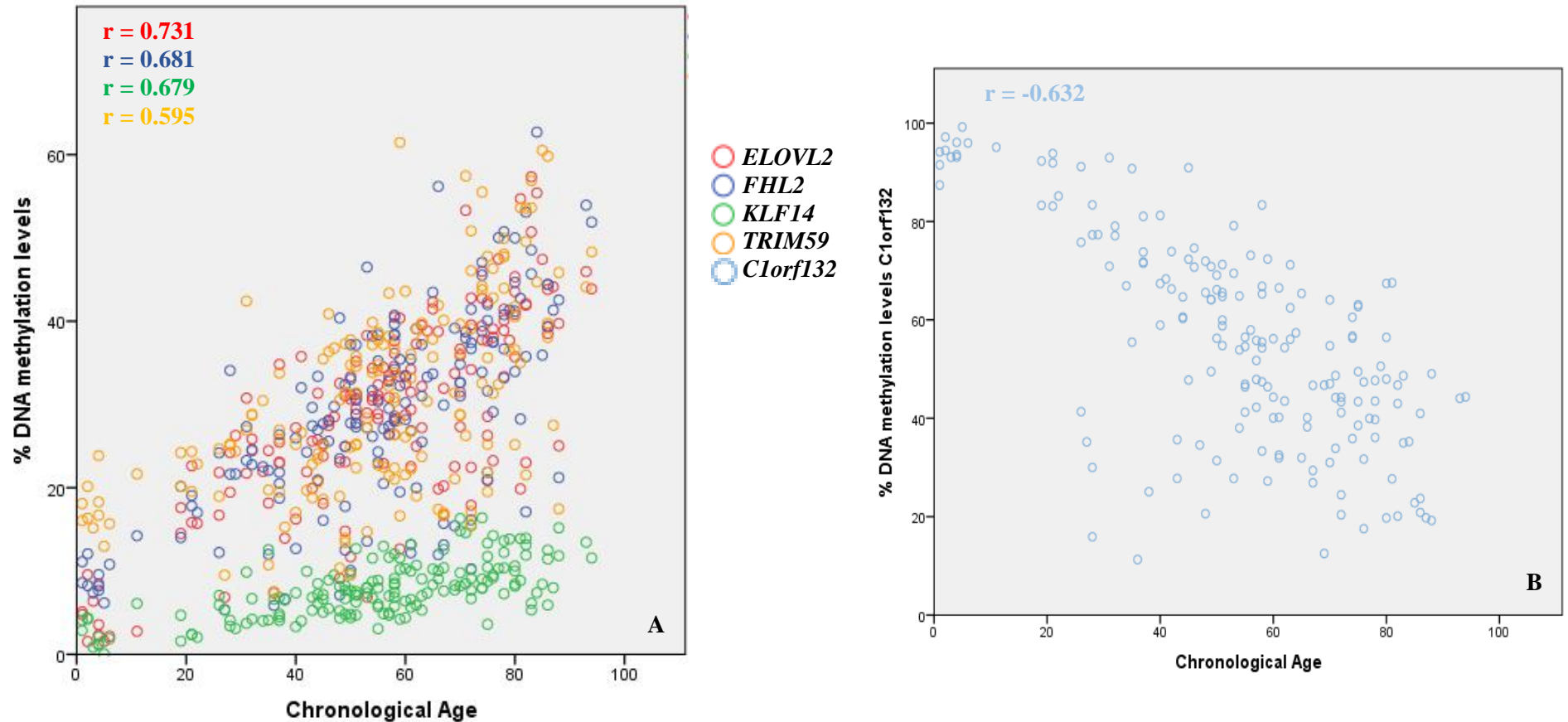
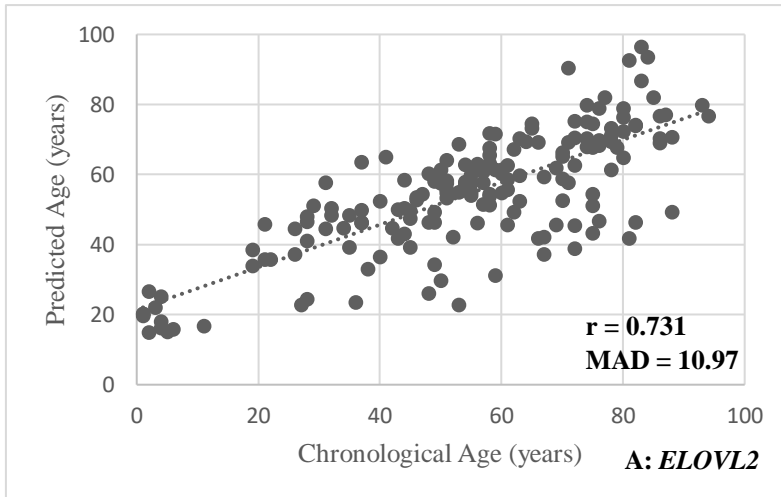
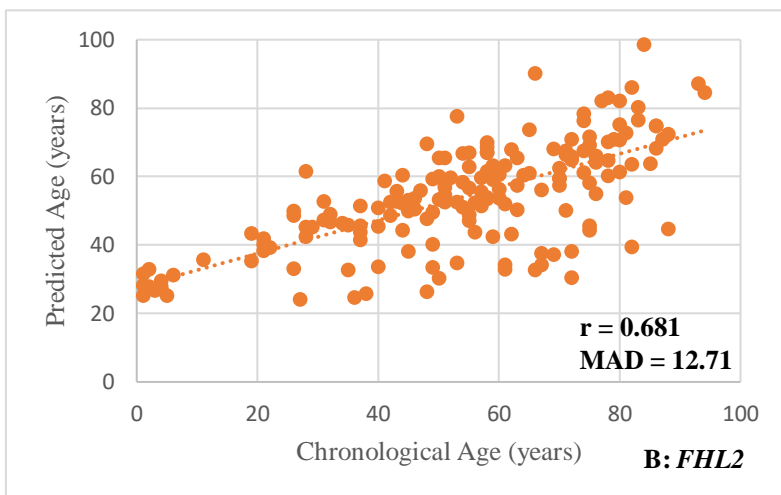


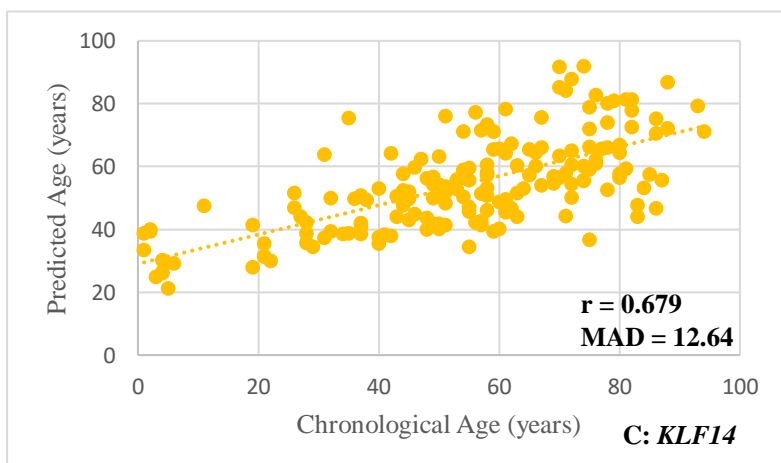
Figure S38: Correlations between DNAm levels and chronological age in 176 samples, including blood samples from living and deceased individuals, bone samples collected from autopsies and teeth from living and deceased individuals, obtained through SNaPshot methodology. A) Positive correlation between methylation levels and chronological age for CpG sites in *ELOVL2* (red), *FHL2* (blue), *KLF14* (green) and *TRIM59* (yellow) genes; B) negative correlation between methylation levels and chronological age for CpG site in *C1orf132* gene (light blue). The corresponding Spearman correlation coefficients (r) are depicted inside each plot.



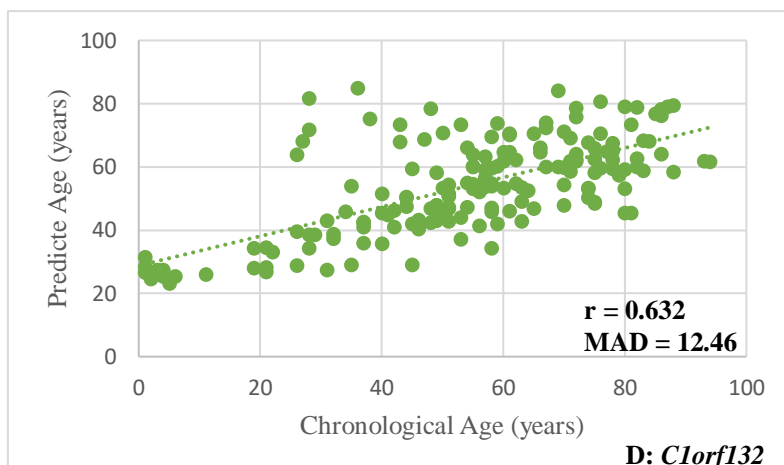
A) $Predicted\ Age = 12.676 + 145.856 \times DNAm\ level\ ELOVL2.$



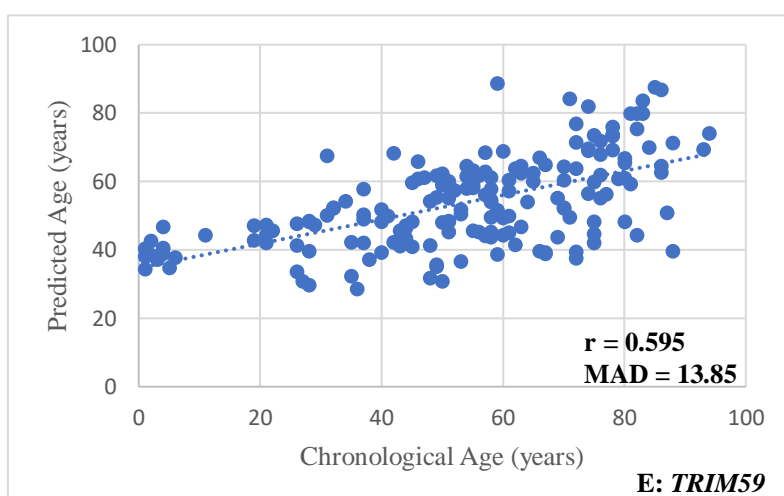
B) $Predicted\ Age = 17.131 + 129.869 \times DNAm\ level\ FHL2.$



C) $Predicted\ Age = 21.152 + 432.557 \times DNAm\ level\ KLF14.$



$$D) \text{ Predicted Age} = 92.997 - 70.410 \times \text{DNAm level } C1orf132.$$



$$E) \text{ Predicted Age} = 20.193 + 111.242 \times \text{DNAm level } TRIM59.$$

Figure S39: Plots with predicted age (years) *versus* chronological age (years) of the 176 samples, including blood samples from living and deceased individuals, tooth samples from living and deceased individuals and bone samples collected from autopsies, using simple linear regression models developed with the five CpGs through the multiplex methylation SNaPshot methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 172 individuals based on methylation levels of the *ELOVL2*; B) Predicted age of the 175 individuals based on methylation levels of the *FHL2*; C) Predicted age of the 174 individuals based on methylation levels of the *C1orf132*; D) Predicted age of the 171 individuals based on methylation levels of the *KLF14*; E) Predicted age of the 175 individuals based on methylation levels of the *TRIM59*.

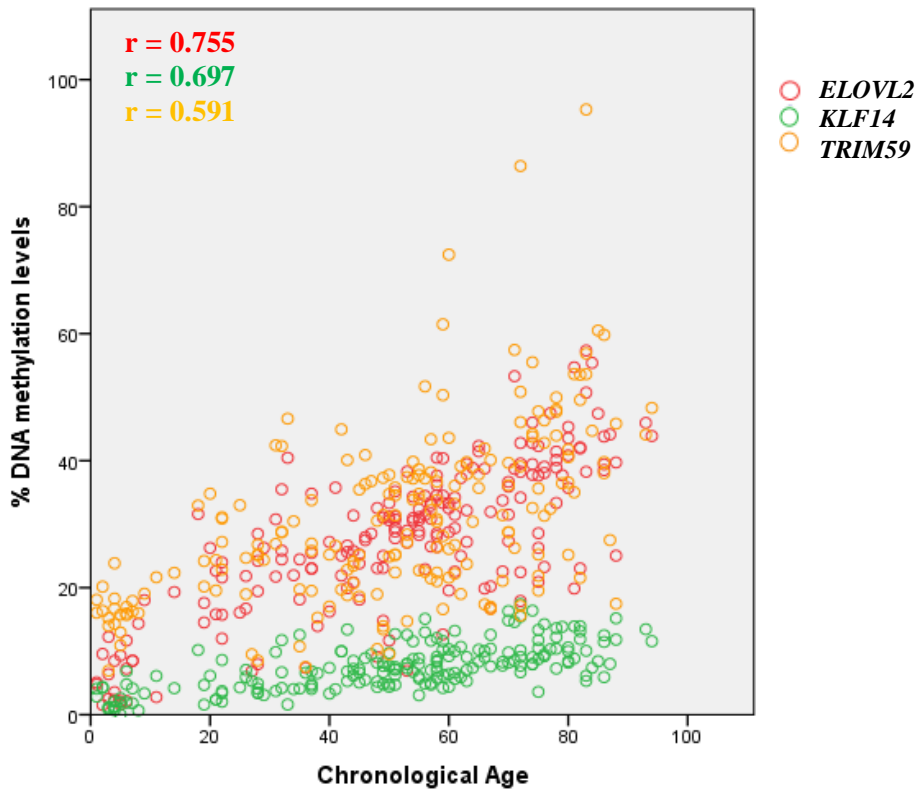
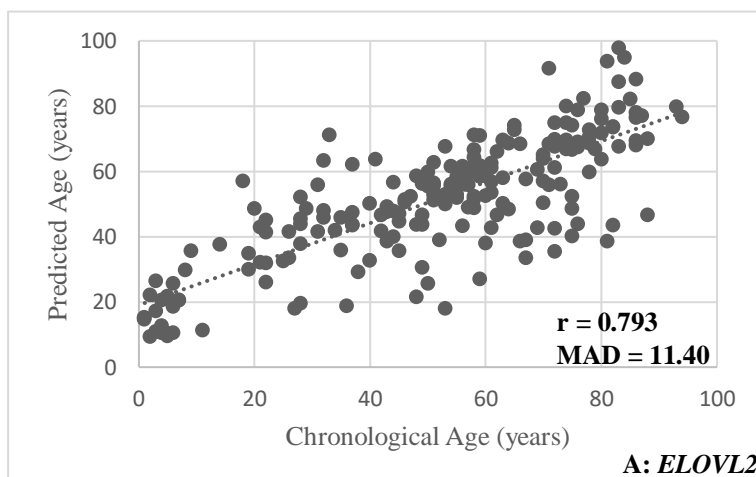
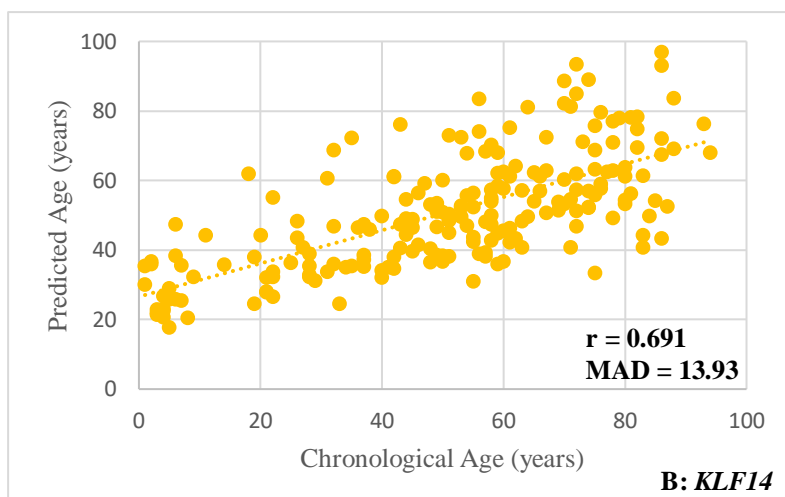


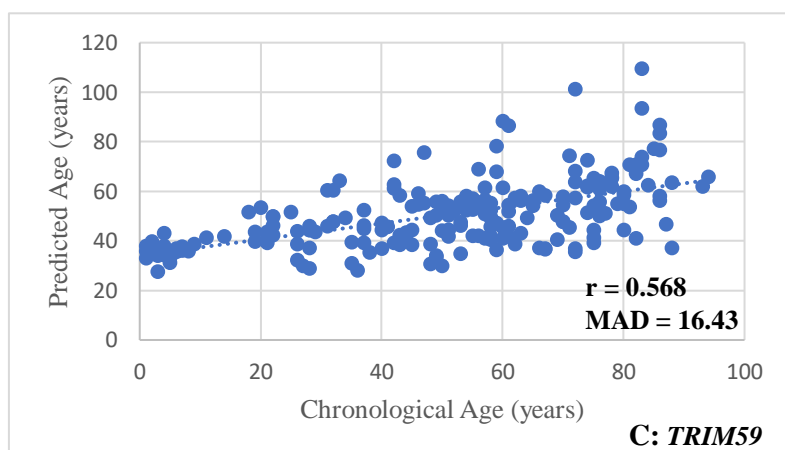
Figure S40: Positive correlation between DNAm levels and chronological age in 215 samples (including blood samples from living and deceased individuals, bone samples collected from autopsies, teeth from living and deceased individuals and buccal swabs from living individuals) obtained through SNaPshot methodology for CpGs in *ELOVL2* (red), *KLF14* (green) and *TRIM59* (yellow) genes. The corresponding Spearman correlation coefficients (r) are depicted inside the plot.



A) $Predicted\ Age = 7.086 + 158.522 \times DNAm\ level\ ELOVL2.$



$$B) \text{ Predicted Age} = 17.651 + 435.565 \times \text{DNAm level } KLF14.$$



$$C) \text{ Predicted Age} = 20.986 + 92.832 \times \text{DNAm level } TRIM59.$$

Figure S41: Plots with predicted age (years) *versus* chronological age (years) of the 215 individuals, including blood samples from living and deceased individuals, bone samples collected from autopsies, teeth from living and deceased individuals and buccal swabs from living individuals, using simple linear regression models developed with the three CpGs in each gene through the multiplex methylation SNaPshot methodology. MAD value and Spearman correlation coefficient (r) are plotted in each chart. A) Predicted age of the 208 individuals based on methylation levels of the CpG from *ELOVL2* gene; B) Predicted age of the 206 individuals based on methylation levels of the CpG from *KLF14* gene; C) Predicted age of the 210 individuals based on methylation levels of the CpG from *TRIM59* gene.

Annex VII

Description of bone samples collected from CEI/XXI (in Portuguese)



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 21/07/2018

Esqueleto: CEI/XXI_87

Conservação geral do esqueleto: Esqueleto frágil, em mau estado de conservação. Fémures no geral degradados, essencialmente na zona das epífises e diáfises.

Fémur esquerdo: Cabeça do fémur em mau estado de preservação, exposição do tecido trabecular na região próxima do colo anatómico; grande Trocânter: exposição do tecido trabecular; pequeno Trocânter: exposição do tecido trabecular; perda do côndilo lateral na região posterior e lateral; evidência de eburnação no côndilo medial; raspagem de osso em meio da diáfise (norma posterior, possível extracção de DNA anterior); rebordo de osteofitose no côndilo medial.

Fémur direito: Cabeça do fémur: exposição do tecido trabecular; grande Trocânter: exposição mais evidenciada do tecido trabecular comparativamente ao fémur esquerdo; pequeno Trocânter: exposição do tecido trabecular; degradação do côndilo lateral (menos severa em comparação ao fémur esquerdo); degradação na região posterior do osso.

Dados AM:

Sexo: F

Idade Cronológica: 83 anos

Ancestralidade:-

Estatura: -

Peça óssea retirada e lateralidade: Fémur esquerdo

Medidas

Não aconselhável (devido a degradação dos ossos) mas retiradas medidas.

Fémur esquerdo

- Comprimento máximo : 36.8 cm
- Comprimento fisiológico: 36.8 cm

Fémur direito

- Comprimento máximo : 37.3 cm
- Comprimento fisiológico: 37.1 cm

Observações:

Fémur esquerdo e fémur direito: presentes fragmentados.

Foto
Norma anterior
Norma posterior
Fotos ao nível dos côndilos
em norma posterior

Fotos



Figura 1: Fémur direito e esquerdo do esqueleto CEI/XXI_87. Fotografia em norma anterior.



Figura 2: Fémur direito e esquerdo do esqueleto CEI/XXI_87. Fotografia em norma posterior. Visível raspagem prévia do osso para possivelmente análises de DNA no fémur esquerdo.



Figura 3: Fémur direito do esqueleto CEI/XXI_87. Fotografia em norma posterior.



Figura 4: Epífise distal do fémur direito do esqueleto CEI/XXI_87. Fotografia norma posterior.



Figura 5: Fémur esquerdo do esqueleto CEI/XXI_87. Fotografica em norma posterior (A e B). Em B evidência de eburnação no cômulo medial.



Figura 6: Epífise distal do fémur esquerdo do CEI/XXI_87. Fotografia em norma posterior, visão lateral.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 21/07/2018

Esqueleto: CEI/XXI_128

Conservação geral do esqueleto: bom estado de conservação

Dados AM:

Sexo: F

Idade Cronológica: 38 anos

Ancestralidade:-

Estatura: -

Peça óssea retirada e lateralidade: Fémur direito

Medidas

Fémur direito

-Comprimento máximo: 44.7 cm

-Comprimento fisiológico: 44.3 cm

Fémur esquerdo

-Comprimento máximo: 44.4 cm

-Comprimento fisiológico: 44.3 cm

Foto
Norma anterior
Norma posterior

Fotos



Figura 1: Fémur direito do esqueleto CEI/XXI_128. Fotografia em norma anterior.



Figura 2: Fémur direito do esqueleto CEI/XXI_128. Fotografia em norma posterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 22/05/2018

Esqueleto: CEI/XXI_296

Conservação geral do esqueleto: bom estado de conservação

Dados AM:

Sexo: M

Idade Cronológica 77 anos

Ancestralidade:-

Estatura: -

Peça óssea retirada e lateralidade: Fémur esquerdo

Medidas

Fémur esquerdo

-Comprimento máximo: 42.6 cm

-Comprimento fisiológico: 42.5 cm

Observações

O indivíduo poderá ter cancro (dados *antemortem*). Não incluir no estudo.

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur direito do esqueleto CEI/XXI_296. Fotografia em norma anterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_ Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_62

Conservação geral do esqueleto:

Fémures preservados mas com alguma escamação essencialmente no fémur direito, côncilo lateral. Recolha de DNA anterior (raspagem) no fémur esquerdo. Ossos pesados.

Dados AM:

Sexo: M

Idade Cronológica: 60 anos

Ancestralidade:-

Estatura: -

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur esquerdo

-Comprimento fisiológico: 42 cm

-Comprimento máximo: 42.2 cm

Fémur direito

-Comprimento fisiológico: 42 cm

-Comprimento máximo: 42.5 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur esquerdo do esqueleto CEI/XXI_ 62. Fotografia em norma anterior.



Figura 2: Fémur esquerdo do esqueleto CEI/XXI_ 62. Fotografia em norma posterior.



Figura 3: Fémur direito do esqueleto CEI/XXI_ 62. Fotografia em norma anterior.



Figura 4: Fémur direito do esqueleto CEI/XXI_ 62. Fotografia em norma posterior.



Figura 5: Fémur direito esqueleto CEI/XXI_ 62. Vísivel erosão, fragmentação óssea no côndilo lateral.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_78

Dados AM:

Sexo: M

Idade Cronológica: 38 anos

Ancestralidade:-

Estatura:-

Conservação: fémures pesados, preservados; recolha de DNA anterior (raspagem) no fémur esquerdo. Fémures ligeiramente fragmentados nas epífises, mas são ossos completos.

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur direito

-Comprimento fisiológico: 41.2 cm

-Comprimento máximo: 41.7 cm

Fémur esquerdo

-Comprimento fisiológico: 41.9 cm

-Comprimento máximo: 42 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur direito do esqueleto CEI/XXI_ 78. Fotografia em norma anterior.



Figura 2: Fémur direito do esqueleto CEI/XXI_ 78. Fotografia em norma posterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_103

Dados AM:

Sexo: F

Idade Cronológica: 92 anos

Ancestralidade:-

Estatura:-

Conservação: fémures leves e fragmentados, côndilo lateral do fémur direito e do fémur esquerdo fragmentados. Côndilo lateral do fémur esquerdo mais fragmentado comparativamente ao direito. Fémur esquerdo, grande trocanter fragmentado. Ossos leves.

Peça óssea retirada e lateralidade: Fémur esquerdo

Medidas: Osso fragmentados; não fazer medições.

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur esquerdo do esqueleto CEI/XXI_103. Fotografia em norma anterior.



Figura 2: Fémur esquerdo do esqueleto CEI/XXI_103. Fotografia em norma posterior.



Figura 3: Fémur esquerdo do esqueleto CEI/XXI_103. Fotografia em norma posterior. Visível fragmentação no côndilo lateral.



Figura 4: Fémur esquerdo do esqueleto CEI/XXI_103. Fotografia em norma anterior. Visível na região superior, o grande trocânter fragmentado.



Figura 5: Fémur direito do esqueleto CEI/XXI_103. Fotografia em norma anterior.



Figura 6: Fémur direito do esqueleto CEI/XXI_103. Fotografia em norma posterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_ Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_122

Dados AM:

Sexo: M

Idade Cronológica: 55 anos

Ancestralidade:-

Estatura:-

Conservação: fémures pesados, com alguma degradação nos cêndilos e na cabeça femural, quer no fémur esquerdo quer no fémur direito.

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur esquerdo

-Comprimento fisiológico: 42 cm

-Comprimento máximo: 42.3 cm

Fémur direito

-Comprimento fisiológico: 42.7 cm

-Comprimento máximo: 42.2 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur direito do esqueleto CEI/XXI_122. Fotografia em norma anterior.



Figura 2: Fémur direito do esqueleto CEI/XXI_122. Fotografia em norma posterior.



Figura 3: Fémur esquerdo do esqueleto CEI/XXI_122. Fotografia em norma anterior.



Figura 4: Fémur esquerdo do esqueleto CEI/XXI_122. Fotografia em norma anterior. Visível fragmentação ossea da cabeça do fémur.



Figura 5: Fémur esquerdo do esqueleto CEI/XXI_122. Fotografia em norma posterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_123

Dados AM:

Sexo: F

Idade Cronológica: 83 anos

Ancestralidade:-

Estatura:-

Conservação: fémures preservados (ossos completos), mas com alguma escamação e fragmentação nas epífises.

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur direito

-Comprimento fisiológico: 39.7 cm

-Comprimento máximo: 42 cm

Fémur esquerdo

-Comprimento fisiológico: 42 cm

-Comprimento máximo: 42.2 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur esquerdo do esqueleto CEI/XXI_123. Fotografia em norma anterior.



Figura 2: Fémur esquerdo do esqueleto: CEI/XXI_123. Fotografia em norma posterior.



Figura 3: Fémur direito do esqueleto CEI/XXI_123. Fotografia em norma anterior.



Figura 4: Fémur direito do esqueleto CEI/XXI_123. Fotografia em norma posterior.



Figura 5: Fémur direito do esqueleto CEI/XXI_123, norma posterior. Visível a erosão óssea no côndilo.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 11/12/2018

Esqueleto: CEI/XXI_267

Dados AM:

Sexo: F

Idade Cronológica: 59 anos

Ancestralidade:-

Estatura:-

Conservação: Estado geral dos fémures é bom (ossos completos) mas com alguma fragmentação, escamação; fémur direito fragmentado no côndilo medial; fémur esquerdo fragmentação no côndilo medial e lateral.

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur direito

Comprimento fisiológico: 37.3 cm

Comprimento máximo: 37.7 cm

Fémur esquerdo

Comprimento fisiológico: 37.5 cm

Comprimento máximo: 37.6 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur esquerdo do esqueleto CEI/XXI_267. Fotografia em norma anterior.



Figura 2: Fémur esquerdo do esqueleto CEI/XXI_267. Fotografia em norma posterior.



Figura 3: Fémur direito do esqueleto CEI/XXI_267. Fotografia em norma anterior.



Figura 4: Fémur direito do esqueleto CEI/XXI_267. Fotografia em norma posterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 8/1/2019

Esqueleto: CEI/XXI_164

Dados AM:

Sexo:-

Idade Cronológica: 48 anos

Ancestralidade:-

Estatura:-

Conservação: Fémures completos mas com alguma fragmentação e escamação óssea nos cêndilos; ossos leves.

Exposição trabecular nos cêndilos do fémur direito na região posterior.

Fémur esquerdo com erosão na cabeça femoral.

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur esquerdo

-Comprimento fisiológico: 38.8 cm

-Comprimento máximo: 39.4 cm

Fémur direito

-Comprimento fisiológico: 38.8 cm

-Comprimento máximo: 39.3 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur direito e esquerdo do esqueleto CEI/XXI_164. Fotografia em norma anterior.



Figura 2: Côndilos lateral e medial do fémur direito do esqueleto CEI/XXI_164. Fotografia em norma posterior. Fragmentação óssea visível, com exposição do osso trabecular.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_ Metilação de DNA

Data: 8/1/2019

Esqueleto: CEI/XXI_178

Dados AM:

Sexo: F

Idade Cronológica: 83 anos

Ancestralidade:-

Estatura:-

Conservação: Bom estado de preservação. Evidência de eburnação nos côndilos femurais.

Peça óssea retirada e lateralidade: -

Lateralidade: -

Medidas: -

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos



Figura 1: Fémur direito e esquerdo do esqueleto CEI/XXI_178. Fotografia em norma anterior.



Figura 2: Eburnação nos côndilos femurais do esqueleto CEI/XXI_178. Fotografia em norma anterior.



Investigadora: **Maria Helena Correia Dias**
Estudante de Doutoramento em Antropologia:
Antropologia Forense
Contacto telefónico: 962903987
Email: helenacorreiadias30@gmail.com

Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Colecção de Esqueletos Identificados do Século XXI

Inventário _HelenaCorreiaDias_Metilação de DNA

Data: 8/1/2019

Esqueleto: CEI/XXI_262

Dados AM:

Sexo: -

Idade Cronológica: 58 anos

Ancestralidade:-

Estatura:-

Conservação: Bom estado de preservação do esqueleto. Fémures completos.

Evidência de alguma erosão óssea nos côndilos dos fémures direito e esquerdo,

Peça óssea retirada e lateralidade: Fémur direito

Medidas:

Fémur esquerdo

-Comprimento fisiológico: 39.6 cm

-Comprimento máximo: 39.8 cm

Fémur direito

-Comprimento fisiológico: 39.7 cm

-Comprimento máximo): 40.1 cm

Foto
Norma anterior
Norma posterior
Norma lateral
Norma medial

Fotos




Figura 1: Fémur direito e esquerdo do esqueleto CEI/XXI_262. Fotografia em norma anterior.

Annex VIII

Published papers

PAPER

CRIMINALISTICS

Helena Correia Dias,^{1,2,3} M.Sc.; Cristina Cordeiro,^{3,4} M.Sc., M.D.; Francisco Corte Real,^{3,4} M.D., Ph.D.; Eugénia Cunha,^{2,3} Ph.D.; and Licínio Manco ¹ Ph.D.

Age Estimation Based on DNA Methylation Using Blood Samples From Deceased Individuals[†]

ABSTRACT: Age estimation using DNA methylation levels has been widely investigated in recent years because of its potential application in forensic genetics. The main aim of this study was to develop an age predictor model (APM) for blood samples of deceased individuals based in five age-correlated genes. Fifty-one samples were analyzed through the bisulfite polymerase chain reaction (PCR) sequencing method for DNA methylation evaluation in genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, and *C1orf132*. Linear regression was used to analyze relationships between methylation levels and age. The model using the highest age-correlated CpG from each locus revealed a correlation coefficient of 0.888, explaining 76.3% of age variation, with a mean absolute deviation from the chronological age (MAD) of 6.08 years. The model was validated in an independent test set of 19 samples producing a MAD of 8.84 years. The developed APM seems to be informative and could have potential application in forensic analysis.

KEYWORDS: forensic science, forensic epigenetics, DNA methylation age, deceased individuals, bisulfite PCR sequencing

Age estimation based on DNA methylation has been widely investigated in recent years. Prediction of age from biological evidences can be very useful in forensic genetics for identification purposes of human remains from mass disasters or to solve crimes by limiting the search range of unknown suspects (1,2).

Several DNA methylation markers have been investigated in various tissues and body fluids using DNA-based methodologies such as bisulfite pyrosequencing (3–9), EpiTYPER technology (10), massively parallel sequencing (11–13), or SNaPshot assays (14–16). This allowed the identification of many CpG markers showing high correlations with chronological age, potentially useful as forensic age predictors. Thus, a number of high accurate age prediction models have been proposed for specific tissues, including blood (6,7,12), teeth (17), buccal swabs (8) or saliva (15), or as multi-tissue models (13,16,18,19).

Most of DNA methylation markers have been investigated and validated mainly in whole blood of living individuals using bisulfite pyrosequencing. Because DNA methylation at genes *ELOVL2*,

FHL2, *EDARADD*, *PDE4C*, and *C1orf132* have been repeatedly reported in independent studies to have strong age association in blood, they are considered to be some of the most promising age-predictive markers for blood samples. Both *ELOVL2* and *EDARADD* genes were included in a shortlist of 44 genomic regions most significantly associated with age in a meta-analysis (20). Also, the *PDE4C* locus was ranked in the three best markers among 102 age-related CpG sites in blood (3). Zbiéc-Piekarska et al. (6) published an assay which included in the overall model 5 CpG sites in the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132/MIR29B2C*, and *TRIM59* genes, resulting in a mean absolute deviation from the chronological age (MAD) of 3.9 years. Cho et al. (9) replicated the strong age association for DNA methylation markers in the *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14*, and *FHL2* genes in blood samples of 100 Korean individuals.

To the best of our knowledge, only three studies have used blood samples from deceased individuals to address the correlation between DNA methylation and chronological age. The study of Bekaert et al. (4), using the pyrosequencing methodology, investigated CpG sites from 4 genes (*ASPA*, *PDE4C*, *ELOVL2*, and *EDARADD*) in blood samples from 169 deceased and 37 living individuals. Naeue et al. (13) investigated 13 CpGs (located in genes *DDO*, *ELOVL2*, *F5*, *GRM2*, *HOXC4*, *KLF14*, *LDB2*, *MEIS1-AS3*, *NKIRAS2*, *RPA2*, *SAMD10*, *TRIM59*, and *ZYG11A*) in blood samples from 29 deceased individuals by massively parallel sequencing. These markers were previously selected as strong age-dependent loci on whole blood from living individuals (12). Hamano et al. (21) analyzed the methylation status of the *ELOVL2* and *FHL2* promoter regions by methylation-sensitive high-resolution melting (MS-HRM) using 22 living and 52 dead blood samples.

¹Department of Life Sciences, Research Centre for Anthropology and Health (CIAS), University of Coimbra, Coimbra, Portugal.

²Department of Life Sciences, Laboratory of Forensic Anthropology, Centre for Functional Ecology (CEF), University of Coimbra, Coimbra, Portugal.

³National Institute of Legal Medicine and Forensic Sciences, Coimbra, Portugal.

⁴Faculty of Medicine, University of Coimbra, Coimbra, Portugal.

Corresponding author: Licínio Manco, Ph.D. E-mail: lmanco@antrop.uc.pt

[†]Supported in part by Fundação para a Ciência e a Tecnologia (FCT) (FCT-PEst-OE/SADG/UI0283/2019). H.D. has a PhD grant from FCT (SFRH/BD/117022/2016).

Received 8 May 2019; and in revised form 9 Aug. 2019; accepted 14 Aug. 2019.

Considering the scarcity of data addressing methylation status in deceased individuals for age estimation purposes, the main aim of the present study was to develop an age predictor model (APM) for blood samples of deceased individuals based on five previously known and validated age-associated genes *ELOVL2*, *EDARADD*, *PDE4C*, *FHL2*, and *C1orf132*.

We report on the labor- and cost-effective semi-quantitative bisulfite polymerase chain reaction (PCR) sequencing method, using the peak height information obtained from chromatograms. Although other groups have developed algorithms for quantitative methylation analysis to help signal normalization and estimation of effectiveness of bisulfite treatment (22), the direct bisulfite PCR sequencing was previously shown to have similar linearity and accuracy to pyrosequencing for methylation evaluation (23,24).

Materials and Methods

Sample Collection

Blood samples from 51 autopsies (aged 24–86 years old; 7 females and 44 males) were collected in *Unidade de Patologia Forense da Delegação do Centro do Instituto Nacional de Medicina Legal e Ciências Forenses (INMLCF)*, after consulting *RENDA (Registo Nacional de Não Dadores)*. Additionally, a set of 19 blood samples (aged 37–88 years; 8 females and 11 males) were collected for validation purposes in *Unidade de Patologia Forense da Delegação do Centro do INMLCF* and from *Bodies Donated to Science (BDS)* before the embalming method with Thiel (25), in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*. All dead bodies had no evidence of cancer, which could affect the methylation status (26). Blood samples were collected within 5 days after death.

The study protocol was approved by the *Instituto Nacional de Medicina Legal e Ciências Forenses* and by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (no 038-CE-2017).

DNA Extraction, Bisulfite Conversion PCR and Sanger Sequencing

Blood samples were collected in EDTA tubes and submitted to DNA extraction using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). DNA extracts were quantified in a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA). Genomic DNA was subjected to bisulfite conversion using EZ DNA Methylation Gold Kit (Zymo Research, Irvine) according to the instructions of manufacturer. Briefly, 20 μL of genomic DNA (in a total amount of 200 to 400 ng) was treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 μL . After conversion, modified DNA samples were submitted to PCR for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, and *C1orf132* using the Qiagen Multiplex PCR kit (Qiagen) using sets of primers previously described (4,6) (Table S1). Total PCR volume was 25 μL , containing 2 μL of converted DNA, 0.2 μM of each primer, and 1 \times Qiagen Multiplex PCR Master Mix (Qiagen). PCR amplification program consisted of an initial step at 95°C for 15 min followed by 40 cycles of 30 sec at 95°C, 1 min at 60°C, and 1 min at 72°C for *ELOVL2* gene and 35 cycles of 30 sec at 95°C, 1 min at 56°C, and 1 min at 72°C for *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes. A final extension of 72°C for 10 min ended

PCR amplification. A negative PCR control was included in each amplification. The size and quality of PCR products were visualized under UV light on 2% agarose gels. A strong band with the anticipated product size should be present for sequencing. Subsequently, the PCR products were purified using ExoSAP-IT (Affymetrix, Cleveland) and sequenced using the reverse primer by the Sanger's dideoxy chain termination reaction, using Big-Dye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems, Foster City) and the ABI 3130 sequencer using the POP-7™ polymer as separation matrix (Applied Biosystems).

Methylation Quantification of DNA Sequencing Data

The methylation status of cytosines (C) in each CpG dinucleotide was estimated by measuring the ratio between peak height values of cytosine (C) and thymine (T) through the formula $[C/C+T]$ in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium, Australia) (Fig. S1). In each CpG, a single C was considered as completely methylated (100%), a single T as completely unmethylated (0%) and overlapping C and T reveal partial methylation (0%–100%). As in our data sequencing with the reverse primer results in a cleaner chromatogram, we used the reverse complement strand in Chromas to estimate the ratio between peak heights of C and T (Fig. S2).

DNA Methylation Standards

Each primer set used for bisulfite sequencing was independently verified, to confirm the accuracy of sequencing data using the DNA methylation commercial standards EpiTect Control DNA, methylated, and unmethylated (Qiagen). Standard DNA samples premixed at methylation levels of 0%, 50%, and 100% were used for bisulfite PCR sequencing analysis (Fig. S3).

Statistical Analyses

The statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (SPSS, IBM Corporation, Armonk, NY). Simple linear regressions were used to analyze relationships between methylation levels and chronological age at each single CpG site. Using the regression coefficients from the highest age-correlated CpG sites from each gene, we predicted age of individuals for each individual gene. The same five highest age-correlated CpG sites were selected for simultaneous analysis using multiple linear regression to build a final age prediction model (APM). The mean absolute deviation between chronological and predicted ages (MAD) and the root mean square error (RMSE) were calculated. For our training set of 51 deceased individuals each obtained MAD value was interpreted as either correct or incorrect using a cutoff value according to the standard error of estimate calculated for the APM (± 8 years). The normality of the dependent variable chronological age was assessed by Shapiro–Wilk and Kolmogorov–Smirnov tests. Multicollinearity was investigated by estimating Spearman correlation coefficients between predictor variables; coefficient values $>70\%$ were considered as signal of multicollinearity. Cook's distance was used to find influential outliers in all the predictor variables. Validation of the APM was performed in an independent set of 19 blood samples from deceased individuals and by 4-fold cross-validation using the training set of 51 individuals. The 4-fold cross-validation consisted in

removing randomly a set of samples from the training set of 51 individuals and to develop 4 independent multiple linear regressions on the remaining samples. The removed samples were used for validation purposes. The MAD values were obtained for each of the 4 independent multiple linear regressions, and the mean value was calculated.

Results

We evaluated the association between chronological age and DNA methylation levels in 51 blood samples from deceased individuals, aged 24–86 years old. Five genes selected based on previous studies were investigated using the bisulfite PCR sequencing methodology. Forty-three CpG sites (*ELOVL2*, 9 CpGs; *EDARADD*, 4 CpGs; *FHL2*, 12 CpGs; *PDE4C*, 12 CpGs; and *C1orf132*, 6 CpGs) were selected for methylation evaluation. The focused CpG positions were those selected in previous studies, with some adjacent CpG sites also considered in the analysis.

The efficiency of bisulfite conversion was confirmed measuring the conversion of a random number of cytosines at non-CpG positions. A mean conversion efficiency of 99.99% was observed over all samples and loci.

To obtain a first overview of the markers and the change of DNA methylation per year, a linear regression for individual CpG sites was performed in the training set of 51 individuals, after addressing the normal distribution of the dependent variable (Table S2). The *ELOVL2* locus showed highly significant values for all the selected CpG sites ($R > 0.66$) reflecting the similar strength of the change in DNA methylation with age across all CpGs. For the remaining four loci, several CpG sites revealed no age dependency (nonsignificant p -values) and some CpGs, although significantly associated, revealed moderate or low correlation values. For *FHL2*, all selected CpGs showed lower or moderate change of DNA methylation with age ($R < 0.50$). Fig. S4 (A, B, C, D, and E) illustrate the correlation between age and DNA methylation levels for the most significant age-associated CpG site in each locus, *ELOVL2* CpG4, *PDE4C* CpG2, *C1orf132* CpG1, *EDARADD* CpG3, and *FHL2* CpG2, and Table 1 shows the linear regression statistics for the five CpG sites. The *ELOVL2* CpG4 showed the strongest correlation with age ($R = 0.785$; $p = 2.39 \times 10^{-11}$) explaining 60.8% of the age variation, followed by *C1orf132* CpG1 ($R = 0.634$; $p = 0.000001$), explaining 38.9% of age variation. For the remaining genes, the strongest age-correlated sites were as follows: *PDE4C* CpG2 ($R = 0.592$; $p = 0.000008$), explaining 33.6% of age variation; *EDARADD* CpG3 ($R = 0.621$; $p = 0.000001$), explaining 37.3% of age variation; and *FHL2* CpG2 ($R = 0.465$; $p = 0.00058$), explaining 20% of age variation. Predicting age through a simple linear regression equation for the individual strongest age-associated markers, the

obtained MAD values were 8.89, 9.35, 9.38, 9.93, and 11.40 years, respectively (see Fig. S5: A, B, C, D, and E).

Because DNA methylation for the best CpG markers showed a linear relationship with age, no statistical transformation of variables was made. In spite of this, as a few existing studies refer a quadratic regression for *ELOVL2* (4), we tested the better fit for the relationship between DNA methylation levels and chronological age. The linear regression model showed the best result for *ELOVL2* CpG4 (linear: $R = 0.785$, MAD = 8.89, RMSE = 9.99; quadratic: $R = 0.780$, MAD = 8.48, RMSE = 10.10).

After addressing the absence of multicollinearity between the five most significant age-associated CpG sites, and confirming the absence of outliers (Cook's $D < 1$), we tested a step-wise model by multiple linear regression joining successively the five markers (data not shown). An increase on age-associated statistical values was observed with the successive addition of CpGs and the overall APM using the 5 CpGs showed the highest correlation coefficient ($R = 0.888$), highly significant ($p = 8.17 \times 10^{-13}$), explaining 76.3% of age variation (Table 1). Applying this model, the age prediction for each individual was obtained through the formula: $(20.495) + 77.938 \times \text{methylation level CpG4 } ELOVL2 + 46.879 \times \text{methylation level CpG2 } FHL2 - 70.729 \times \text{methylation level CpG3 } EDARADD + 28.741 \times \text{methylation level CpG2 } PDE4C - 47.984 \times \text{methylation level CpG1 } C1orf132$. Fig. 1A presents a plot with chronological age versus predicted age in the training set of 51 individuals. A strong correlation between predicted and chronological ages was observed (Spearman correlation coefficient, $r = 0.868$), with a MAD of 6.08 years (Table 1). The success rate of correct predictions was 64% assuming that chronological and predicted ages match ± 8 years.

The accuracy of the overall APM was made through a 4-fold cross-validation in the training set, producing a MAD (mean value obtained for the 4 test set) of 7.22 years (RMSE = 7.43). These values were very close to the MAD of 6.08 (RMSE = 7.49) obtained in the whole training set. Additionally, we used an independent set of 19 blood samples from deceased individuals to test the developed APM (Fig. 1B), obtaining a MAD of 8.84 years (RMSE = 10.98).

In addition, to evaluate the accuracy of methylation levels obtained by bisulfite sequencing, we detected the PCR mixture amplification for each locus using three different methylation rates of 0%, 50%, and 100% (Fig. S3). Bisulfite sequencing resulted in DNA methylation levels that bore a significant linear relationship to expected methylation levels.

Discussion

In the present study, we investigated several age-associated CpG markers from five loci (*ELOVL2*, *FHL2*, *EDARADD*,

TABLE 1—Linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes to test for association between CpG sites and chronological age in blood samples of 51 deceased individuals

Locus	CpG Site	Chr. Location	R	R ²	Corrected R ²	SE	p	MAD	RMSE
Simple linear regression									
<i>ELOVL2</i>	CpG4	Chr6: 11044640	0.785	0.617	0.608	10.20	2.39×10^{-11}	8.89	9.99
<i>C1orf132</i>	CpG1	Chr1: 207823681	0.634	0.402	0.389	12.68	0.000001	9.38	12.42
<i>EDARADD</i>	CpG3	Chr1: 236557683	0.621	0.385	0.373	12.72	0.000001	9.93	12.47
<i>PDE4C</i>	CpG2	Chr19: 18343860	0.592	0.350	0.336	13.29	0.000008	9.35	12.75
<i>FHL2</i>	CpG2	Chr2: 105399288	0.465	0.216	0.200	14.36	0.000584	11.40	14.08
Multiple linear regression									
APM			0.888	0.788	0.763	8.02	8.17×10^{-13}	6.08	7.49

Abbreviations: R, correlation coefficient; SE, standard error; MAD, mean absolute deviation; APM, age prediction model; RMSE, root mean square error. Location of CpGs for *ELOVL2*, *FHL2* and *C1orf132* is according to the human GRCh38/hg38 assembly and GRCh37/hg19 for *EDARADD*.

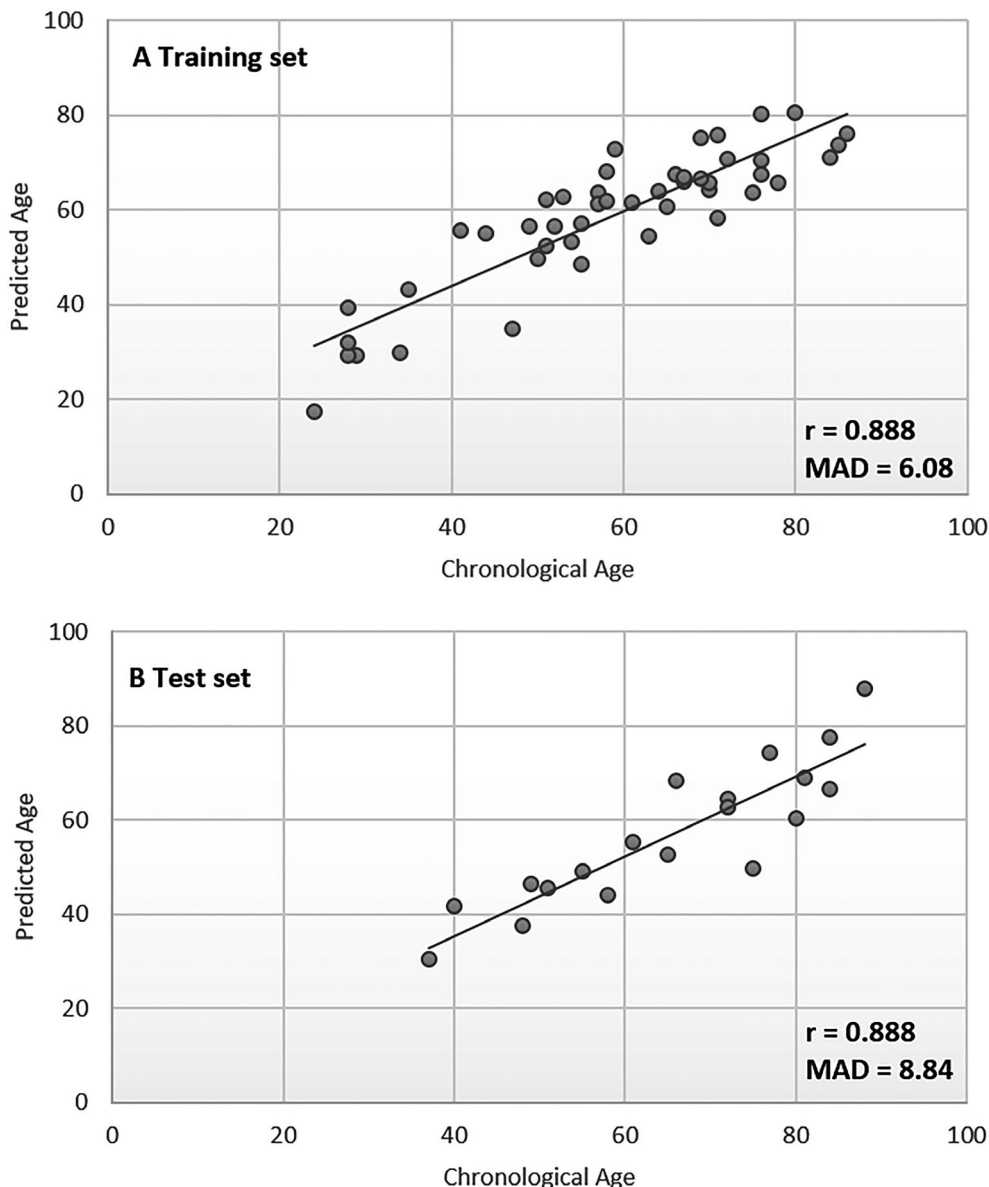


FIG. 1—Plots with predicted versus chronological ages (years) of the training set of 51 individuals (A) and the test set of 19 samples using the final age prediction model (B) based on *ELOVL2* CpG4, *FHL2* CpG2, *EDARADD* CpG3, *PDE4C* CpG2 and *C1orf132* CpG1 markers. The Pearson correlation value (r) was 0.888 in training set (A) as well as in the test set (B). The MAD value was 6.08 years in training set (A) and 8.84 in the test set (B).

PDE4C, and *C1orf132*) in blood samples of deceased individuals using the bisulfite PCR sequencing methodology. Methylation in *ELOVL2*, *FHL2*, and *PDE4C* genes showed a positive correlation with age, whereas methylation in *EDARADD* and *C1orf132* genes was negatively correlated with age. Only the *ELOVL2* gene showed highly significant age-correlation values for all the selected CpG sites reflecting a similar strength on the change in DNA methylation with chronological age. This result is in concordance with previous reports showing *ELOVL2* as the most strong age predictor locus across different tissues (5,13,16,27). The remaining genes *EDARADD*, *FHL2*, *PDE4C*, and *C1orf132* reflect lower correlations between DNA methylation and age, with several CpG sites revealing moderate or absence on age dependency. Nevertheless, the model using the five highest age-associated CpG sites (*ELOVL2* CpG4, *C1orf132* CpG1, *EDARADD* CpG3, *PDE4C* CpG2, and *FHL2* CpG2) showed the best results in age correlation in the training set of

51 individuals, explaining 76.3% of variation in age, and a MAD of 6.08 years. The cross-validation of the training set leads to a MAD of 7.22 years and evaluation of model performance using an independent test set of 19 individuals showed a MAD of 8.84 years. Previous forensic studies, most of them with blood samples of living individuals, using different loci and different number of markers, covering different age ranges and using different methodologies, gave values of MAD between 3.5 to 7.5 years (3,5,7,9,12,16). Thus, in our study the obtained MAD value, although less accurate, is in the range of other research approaches.

In regard to the studies that have used blood samples from deceased individuals addressing the correlation between DNA methylation and age, Bekaert et al. (4), using the pyrosequencing methodology, analyzed 169 blood samples of deceased individuals (including 37 blood samples from living individuals) (age range 0–91 years) for *ELOVL2*, *PDE4C*, *EDARADD*, and *ASPA*

genes. The authors proposed an overall prediction model using the four best age-associated CpG sites from each gene, which was proven to be highly accurate, explaining 95% of the variation in age with a MAD of 3.75 years. Interestingly, the highest age-correlated CpG sites from *ELOVL2* (CpG6; Chr6: 11044640) and *EDARADD* (CpG3; Chr1: 236557683) found in our study correspond to the same best positions observed by Bekaert et al. (4). From *PDE4C*, our best CpG site (CpG2, Chr19: 18343860) was not investigated by Bekaert et al. (4), that found CpG1 (Chr19: 18343888) as the highest age-correlated marker.

Naue et al. (13) investigated through massively parallel sequencing 13 previously selected age-dependent loci in tissues such as brain, bone, muscle, buccal swabs, and whole blood of 29 deceased individuals (age range 0–87 years), including among others the *ELOVL2* locus. In this study of Naue et al. (13), all the analyzed markers in blood (*DDO*, *ELOVL2*, *F5*, *GRM2*, *HOXC4*, *KLF14*, *LDB2*, *NKIRAS2*, *RPA2*, *SAMD10*, *TRIM59*, *MEIS1*, and *ZYG11A*) showed a comparable age dependency in comparison to a previous study using whole blood of living individuals (12). Interestingly, the *ELOVL2* position (Chr6: 11044644; cg16867657) showed a Pearson correlation coefficient of 0.88, which is similar to the value ($r = 0.76$) obtained in our data for the same position (CpG6).

The work of Hamano et al. (21) using the MS-HRM method to address the methylation status of *ELOVL2* and *FHL2* genes revealed similar distributions of DNA methylation levels in 22 living blood samples and 52 dead blood samples. However, the authors suggested that potential differences in methylation status between living and dead samples could be ignorable by the MS-HRM method.

We should note that in our work, differences in accuracy were observed in *ELOVL2* vs. the other markers. The *ELOVL2* marker correlated highly with age and showed only slight differences between CpG sites, in a similar way to the previous studies with whole blood using other methodologies. The obtained differences in DNA methylation accuracy of age markers, with *ELOVL2* as exception, could be explained by modifications on the methylation status of the studied genes by various factors including diseases (28). The possibility that postmortem changes can alter the methylation status among specific loci should also be hypothesized, although this issue has not yet been clarified until now (21). Moreover, it is well known that *ELOVL2* gene is a stable age-associated marker that exhibit consistent age-related changes across different tissues such as whole blood, teeth or saliva (5,13,16,27).

We should also consider that differences in sample size, DNA methylation markers, population age ranges, laboratory methodologies, or statistical techniques could influence accuracies across the different studies. In particular, bisulfite sequencing is a semi-quantitative method and thus may not be optimal for precise methylation analysis. Even though, using DNA methylation standards, the analysis for the best CpG site selected from each gene (*ELOVL2* CpG4, *EDARADD* CpG3, *FHL2* CpG2, *C1orf132* CpG1, and *PDE4C* CpG2) (Fig. S3) indicates that the bisulfite sequencing method is accurate in terms of obtained methylation versus expected methylation of known quantities of methylated to unmethylated DNA.

In conclusion, the present study addressed age dependency of multiple CpG sites in five genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C*, and *C1orf132* through bisulfite Sanger sequencing in blood samples of deceased individuals. We verified that all selected CpG sites in *ELOVL2* showed strong age correlations ($R > 0.66$), as well as three CpGs in *C1orf132* ($R > 0.57$), two in *PDE4C*

($R > 0.58$), and one in *EDARADD* ($R = 0.62$). *FHL2* CpGs showed the lower methylation age dependency. In spite of this, the combination of the five most strong age-correlation markers from each gene in a final APM showed a MAD of 6.08 that seems to be informative and accurate for age prediction purposes.

Acknowledgments

The authors thank all those who supported the present study, including the people who collected blood samples from deceased individuals in INMLCF and also the people who collected blood samples from bodies donated to science in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*.

References

- Jung SE, Shin KJ, Lee HY. DNA methylation-based age prediction from various tissues and body fluids. *BMB Rep* 2017;50(11):546–53.
- Parson W. Age estimation with DNA: from forensic DNA fingerprinting to forensic (Epi) genomics: a mini-review. *Gerontology* 2018;64(4):326–32.
- Weidner C, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol* 2014;15(2):R24.
- Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics* 2015;10(10):922–30.
- Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples. *Forensic Sci Int Genet Suppl Ser* 2015;5:e144–5.
- Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Ż, Pałeczka A, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci Int Genet* 2015;17:173–9.
- Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Ż, Spas A, Parys-Proszek A, et al. Examination of DNA methylation status of the *ELOVL2* marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet* 2015;14:161–7.
- Eipel M, Mayer F, Arent T, Ferreira MRP, Birkhofer C, Costa IG, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. *Aging* 2016;8(5):1034–8.
- Cho S, Jung S-E, Hong SR, Lee EH, Lee JH, Lee SD, et al. Independent validation of DNA-based approaches for age prediction in blood. *Forensic Sci Int Genet* 2017;29:250–6.
- Freire-Aradas A, Phillips C, Mosquera-Miguel A, Girón-Santamaría L, Gómez-Tato A, Casares de Cal M, et al. Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci Int Genet* 2016;24:65–74.
- Aliferi A, Ballard D, Gallidabino MD, Thurtle H, Barron L, Syndercombe Court D. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Sci Int Genet* 2018;37:215–26.
- Naue J, Hoefsloot H CJ, Mook ORF, Rijlaarsdam-Hoekstra L, van der Zwalm MCH, Henneman P, et al. Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. *Forensic Sci Int Genet* 2017;31:19–28.
- Naue J, Sängner T, Hoefsloot H CJ, Lutz-Bonengel S, Kloosterman AD, Verschure PJ. Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. *Forensic Sci Int Genet* 2018;36:152–9.
- Lee HY, Jung S-E, Oh YN, Choi A, Yang WI, Shin K-J. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. *Forensic Sci Int Genet* 2015;19:28–34.
- Hong SR, Jung SE, Lee EH, Shin KJ, Yang WI, Lee HY. DNA methylation based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Forensic Sci Int Genet* 2017;29:118–25.
- Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY. DNA methylation of the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*/*MIR29B2C*, and *TRIM59* genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci Int Genet* 2019;38:1–8.

17. Giuliani C, Cilli E, Bacalini MG, Pirazzini C, Sazzini M, Gruppioni G, et al. Inferring chronological age from DNA methylation patterns of human teeth. *Am J Phys Anthropol* 2016;159(4):585–95.
18. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;14(10):R115.
19. Alsaleha H, McCalluma NA, Halliganb DL, Hadrillia PR. A multi-tissue age prediction model based on DNA methylation analysis. *Forensic Sci Int Genet Suppl Ser* 2017;6:e62–4.
20. Bacalini MG, Boattini A, Gentilini D, Giampieri E, Pirazzini C, Giuliani C, et al. A meta-analysis on age-associated changes in blood DNA methylation: results from an original analysis pipeline for Infinium 450k data. *Aging* 2015;7(2):97–109.
21. Hamano Y, Manabe S, Morimoto C, Fujimoto S, Ozeki M, Tamaki K. Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting. *Leg Med* 2016;21:5–10.
22. Lewin J, Schmitt AO, Adorján P, Hildmann T, Piepenbrock C. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplicates. *Bioinformatics* 2004;20(17):3005–12.
23. Jiang M, Zhang Y, Fei J, Chang X, Fan W, Qian X, et al. Rapid quantification of DNA methylation by measuring relative peak heights in direct bisulfite-PCR sequencing traces. *Lab Invest* 2010;90(2):282–90.
24. Parrish RR, Day JJ, Lubin FD. Direct bisulfite sequencing for examination of DNA methylation patterns with gene and nucleotide resolution from brain tissues. *Curr Protoc Neurosci* 2012;7(7):24.
25. Eisma R, Lamb C, Soames RW. From formalin to thiel embalming: what changes? One anatomy department's experiences *Clin Anat* 2013;26(5):564–71.
26. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013;49(2):359–67.
27. Sliker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin* 2018;11(1):25.
28. Spólnicka M, Pośpiech E, Peplowska B, Zbieć-Piekarska R, Makowska Ż, Pięta A, et al. DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups. *Int J Legal Med* 2018;132(1):1–11.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Example of peaks height determination in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium). In each CpG, the blue arrow (peak height C representing methylated cytosines) and the red arrow (peak height T representing unmethylated cytosines) are compared. The % of DNA methylation level is calculated through the formula $[C/C+T]$. In the present example, we use the *EDARADD*

locus, and we choose a chromatogram of a younger adult individual (female, 28 years).

Figure S2. Examples of sequencing chromatograms for different methylation levels of *ELOVL2*, *EDARADD*, *FHL2*, *PDE4C* and *C1orf132* genes. (A) CpG4 from *ELOVL2* gene (49 years, male); (B) CpG3 from *EDARADD* gene (28 years, female); (C) CpG2 from *FHL2* gene (84 years, male); (D) CpG2 from *PDE4C* gene (69 years, female); (E) CpG1 from *C1orf132* (35 years, male). The blue arrows show the CpG used in the APM.

Figure S3. Verification of bisulfite sequencing for the best CpG site selected from each gene (*ELOVL2* CpG4, *EDARADD* CpG3, *FHL2* CpG2, *C1orf132* CpG1, and *PDE4C* CpG2) using DNA methylation standards (EpiTect Control DNA, methylated and unmethylated - Qiagen). A) Graphic shows the % of methylation obtained in each CpG site for each level of expected methylation (0%, 50% and 100%). Results indicate accuracy of direct bisulfite sequencing across CpGs. B) Graphic shows the obtained methylation vs. the expected methylation using 0%, 50% and 100% DNA methylation standards. Bisulfite sequencing resulted in DNA methylation levels that bore a significant linear relationship to expected methylation levels.

Figure S4. Correlation between methylation levels and chronological age for the most significant age associated CpG sites within each locus, CpG4 of *ELOVL2* (A), CpG1 of *C1orf132* (B), CpG3 of *EDARADD* (C), CpG2 of *PDE4C* (D) and CpG2 of *FHL2* (E) obtained in 51 deceased individuals. The corresponding Spearman correlation coefficients (r) and sample sizes (N) are depicted inside each plot.

Figure S5. Plot with predicted age (years) versus chronological age (years) of the 51 individuals for CpG4 of *ELOVL2* (A), CpG2 of *PDE4C* (B), CpG1 of *C1orf132* (C), CpG3 of *EDARADD* (D) and CpG2 of *FHL2* (E). The corresponding Spearman correlation coefficients (r) are: A) 0.706, B) 0.579, C) 0.649, D) 0.597 and E) 0.469.

Table S1. PCR primers and sequence to analyse for PCR-sequencing.

Table S2. Univariate linear regression analysis of the 43 CpGs sites in *ELOVL2* (9 CpGs), *FHL2* (12 CpGs), *EDARADD* (4 CpGs), *PDE4C* (12 CpGs) and *C1orf132* (6 CpGs) loci in 51 deceased individuals.



DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay



Helena Correia Dias^{a,b,c}, Cristina Cordeiro^{c,d}, Janet Pereira^e, Catarina Pinto^e,
Francisco Corte Real^{c,d}, Eugénia Cunha^{b,c}, Licínio Manco^{a,*}

^a Research Centre for Anthropology and Health (CIAS), Department of Life Sciences, University of Coimbra, Portugal

^b Centre for Functional Ecology (CEF), Laboratory of Forensic Anthropology, Department of Life Sciences, University of Coimbra, Portugal

^c National Institute of Legal Medicine and Forensic Sciences, Portugal

^d Faculty of Medicine, University of Coimbra, Portugal

^e Department of Hematology, Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra, Portugal

ARTICLE INFO

Article history:

Received 27 February 2020

Received in revised form 20 March 2020

Accepted 22 March 2020

Available online 16 April 2020

Keywords:

Methylation SNaPshot

Age prediction

Blood samples

Living and deceased individuals

Replication study

ABSTRACT

Many studies in the forensic field have reported that analysis of DNA methylation is the most reliable method of predicting age. In a previous study, 5 CpG sites located in *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were tested for age prediction purposes in blood, saliva and buccal swab samples from Korean individuals using a multiplex methylation SNaPshot assay. The main goals of the present study were i) to replicate the same multiplex SNaPshot assay in blood samples from Portuguese individuals, ii) to compare DNA methylation status between two different populations and iii) to address putative differences in the methylation status between blood from living and deceased individuals. Blood samples from 59 living individuals (37 females, 22 males; aged 1–94 years-old) and from 62 deceased individuals (13 females, 49 males; aged 28–86 years-old) were evaluated. The specific primers were those previously described. Linear regression models were used to analyse relationships between methylation levels and chronological age using IBM SPSS software v.24. Our results allowed to build a final age prediction model (APM) for blood samples of living individuals with 3 CpG sites, at *ELOVL2*, *FHL2* and *C1orf132* genes, explaining 96.3% of age variation, with a mean absolute deviation (MAD) from chronological age of 4.25 years. Some differences were found in the extent of the age association in the targeted loci comparing Portuguese with Korean individuals. The final APM built for deceased individuals included 4 CpG sites, at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes, explaining 79.3% of age variation, with a MAD of 5.36 years. Combining both sets of samples from living and deceased individuals, the most accurate APM with 4 CpGs, at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes, explained 92.5% of variation in age, with a MAD of 4.97 years. In conclusion, our study replicated in blood samples of Portuguese living individuals a previous SNaPshot assay for age estimation. The possibility that age markers might be population specific and that *postmortem* changes can alter the methylation status among specific loci was suggested by our data. Our study showed the usefulness of the multiplex methylation SNaPshot assay for forensic analysis in blood samples of living and deceased individuals.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Many studies in the field of forensics have reported that analysis of DNA methylation is the most reliable method of predicting age [1,2]. This allowed the identification of many CpG markers in various tissues and body fluids, with high correlations between DNA methylation and chronological age, potential useful in

forensic analysis. Most age-dependent DNA methylation markers have been identified using methodologies for target bisulfite sequencing, including pyrosequencing [3–8], Sanger sequencing [9], EpiTyper [10], or massive parallel sequencing [11–13]. In recent years, the SNaPshot reaction was also introduced as a useful tool for DNA methylation analysis, addressing different tissues and markers including blood, saliva, buccal swabs and semen of living individuals [14–17].

Using the multiplex methylation SNaPshot method, the recent study by Jung et al. [17], tested 5 CpG sites located in *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*/MIR29B2C and *TRIM59* genes for age prediction purposes in blood, saliva and buccal swab samples of healthy

* Corresponding author at: Research Centre for Anthropology and Health (CIAS), Department of Life Sciences, University of Coimbra, 3000-456 Coimbra, Portugal.
E-mail address: lmanco@antrop.uc.pt (L. Manco).

Korean individuals. The age-predictive linear regression model using the 5 CpG sites showed a high correlation between predicted and chronological ages in blood samples, with a Mean Absolute Deviation (MAD) from the chronological age of 3.174 years. The model using these 5 CpG sites was based on the model proposed by Zbieć-Piekarska et al. [5] in blood samples from 120 Polish individuals using pyrosequencing, showing high prediction accuracy with a MAD of 3.9 years. In addition, a MAD of 4.2 years was obtained in another test set of 100 Korean blood samples using the same markers and pyrosequencing methodology [8]. Moreover, at least one or more of these same genes were previously investigated for forensic purposes in blood samples of living [4,6,8] or deceased individuals [4,12,18].

Most studies on DNA methylation for age estimation purposes focused on the identification of new sets of markers in a specific population group or training specific methodologies. However, the validation and replication of experiments to test proposed age-predictive DNA markers and methodologies are strongly recommended for forensic applications to establish consistency between populations and laboratories [8]. Moreover, the possibility that *postmortem* changes could alter the methylation status when performing age prediction should also be considered in forensic cases and, so far, only two studies focused on this point [4,18].

That said, the main goals of the present study were as follows: i) to replicate in blood samples from Portuguese individuals the multiplex SNaPshot assay proposed by Jung et al. [17]; ii) to compare DNA methylation status between two different populations; and iii) to address putative differences in the methylation status between blood samples from living and deceased individuals. We re-analyzed DNA methylation detection using the multiplexing capability of SNaPshot because this methodology could be a simple, efficient and cost-effective way to determine simultaneously DNA methylation levels for different individual target CpG sites in the field of forensics.

2. Material and methods

2.1. Sample collection

Peripheral blood samples from 59 Portuguese healthy living individuals (37 females, 22 males; aged 1–94 years-old) were collected from users of *Bio banco - Hospital Pediátrico de Coimbra*. Written informed consent was previously obtained from adult participants and from children's parents, under the age of 18 years. A total of 62 blood samples from deceased individuals were collected: 55 blood samples (11 females, 44 males; aged 28–86 years) from Portuguese individuals with no known diseases were collected during autopsy in *Serviço de Patologia Forense da Delegação do Centro*, after consulting *RENDA (Registo Nacional de Não Dadores)*; and 7 blood samples (2 females, 5 males; aged 49–84 years) obtained from the Bodies Donated to Science (BDS) were collected before the embalming method with Thiel [19] in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*. All the blood samples were collected within 5 days after death.

The study protocol was approved by the Instituto Nacional de Medicina Legal e Ciências Forenses and by the Ethical Committee of Faculdade de Medicina da Universidade de Coimbra (nº 038-CE-2017).

2.2. DNA extraction, bisulfite conversion and SNaPshot assay

Blood samples from living and deceased individuals collected in EDTA-tubes and stored frozen at -20°C were subjected to DNA extraction using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany).

DNA extracts were quantified in a Nanodrop spectrophotometer (Thermo Fisher Scientific) and subjected to bisulfite conversion using EZ DNA Methylation Gold Kit (Zymo Research, Irvine, USA) according to the instructions of manufacturer. Briefly, 20 μL of genomic DNA (in a total amount of 200–400 ng) were treated with sodium bisulfite and modified DNA was extracted to a final volume of 10 μL .

After bisulfite conversion, the modified DNA samples were submitted to a multiplex polymerase chain reaction (PCR) amplification for the 5 CpG sites at genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* with the primers previously described in Jung et al. [17], using the Qiagen[®] Multiplex PCR kit (Qiagen). Total PCR volume was 12.5 μL , containing 1 μL (40–60 ng) of converted DNA, 0.5 μL of pooled primers (μM concentration of each primer according [17]) and 6.25 μL of 2x Qiagen[®] Multiplex PCR Master Mix (Qiagen). The PCR amplification program consisted of an initial hold at 95°C for 11 min followed by 25 cycles of 20 s at 94°C , 1 min at 56°C , and 30 s at 72°C . A final extension of 72°C for 7 min ended PCR amplification. A negative PCR control was included in each amplification. The size and quality of PCR products were visualized on 2% agarose gels with ethidium bromide under UV light.

Subsequently, the PCR products were purified using 1 μL of amplified products with 0.5 μL of ExoSAP-IT (Affymetrix, Cleveland, USA) in an initial purification reaction (37°C for 15 min followed by 80°C for 15 min). Next, 1.5 μL of the purified PCR products was sequenced using a multiplex SBE (single-base extension) reaction with the addition of 1 μL of Ready Reaction Mix SNaPshot Multiplex Kit (Applied Biosystems, Foster City, USA), 0.5 μL of a pooled primer mix of sequencing primers (μM concentrations according [17]) and 2 μL H_2O (in a total volume of 5 μL). Sequencing program consists in 10 s at 96°C , following 5 s at 50°C and 30 s at 60°C for 25 cycles. A final purification was made with 1 μL of Shrimp Alkaline Phosphatase Recombinant rSAP (Applied Biosystems). The SBE reactions were analysed using the SeqStudio Genetic Analyzer (Applied Biosystems) and the GeneMapper Software 6 (Applied Biosystems).

The methylation levels at each CpG site (0–1) from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were calculated from the nucleotide intensities measured by peak heights observed in the electropherograms, as described in Jung et al. [17].

2.3. Statistical analysis

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (SPSS, IBM Corporation, Armonk, NY). Independent analyses were made for the sets of living, deceased and the overall sample set of blood samples. Simple linear regressions were used to analyze relationships between DNA methylation levels and chronological age at individual CpG sites. Multiple regression analysis were performed with the 5 CpGs from genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* in living, deceased and overall samples in order to select the statistically significant predictor variables to be used in the final age prediction model (APM) developed in each set of samples.

The MAD values between predicted and chronological ages were calculated for the final APM constructed using the regression coefficients from each sample set. Additionally, MAD values were calculated using the final APM model, for subsets of distinct age categories in the training set of living and deceased blood samples. Four groups were considered for living individuals (<18 years, 19–39 years, 40–60 years and >61 years) and three distinct age categories were addressed for deceased individuals (28–51 years, 52–71 years and 72–88 years).

Validation of the APM was performed by a 4-fold cross validation using both individual training sets from living and deceased individuals and the overall data set of blood samples. The

cross validation consisted in removing randomly a set of samples from the data set and assigned as a validation set while the rest of the data is used as training set. An additional validation was performed, in which the complete data set was split into 2 sets (training and validation sets), and an independent regression was calculated for the training set and applied to the validation set.

For the evaluation of differences between sexes, we made a comparison of regression lines relating chronological age and DNA methylation levels of each gene at two levels (males/females) of the categorical factor. Analysis was made to determine if there are significant differences between the slopes and the intercepts at the two levels of that categorical factor, using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA).

3. Results

In the present study, we reanalyzed the multiplex methylation SNaPshot assay of Jung et al. [17] in blood samples from 59 healthy individuals (1–94 years old; 37 females and 22 males) of Portuguese ancestry. Moreover, we analyzed 62 blood samples of deceased individuals (aged 28–86 years old; 54 males and 16 females) to compare methylation data between both groups.

Two simple linear regression lines of methylation status and age between males and females showed no statistically significant difference in slope and intercept in the overall sample set of living and deceased individuals (Supplementary Table S1). Thus, all the analyses were made ignoring gender differences.

DNA methylation levels at 5 CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes were simultaneously measured through a SNaPshot assay (Supplementary Fig. S1), and independent analyses were made for living, deceased and the overall blood samples. In all sets of samples, positive correlations were observed between DNA methylation and chronological age for *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes (Supplementary Fig. S2), and a negative correlation with age was obtained for *C1orf132* (Supplementary Fig. S3).

3.1. Replication of a SNaPshot assay in blood samples from living individuals

In blood samples from living individuals (aged 1–94 years old), all the 5 CpG sites showed strong correlations between DNA

methylation and chronological age (Supplementary Figs. S2A & S3A). The strongest correlation was observed for *ELOVL2* locus ($R=0.951$; $P=3.58 \times 10^{-29}$), followed by *FHL2* ($R=0.946$; $P=1.49 \times 10^{-29}$), *C1orf132* ($R=-0.924$; $P=1.67 \times 10^{-25}$) and *TRIM59* ($R=0.910$; $P=2.04 \times 10^{-23}$). The CpG at *KLF14* gene showed the lowest age correlation value ($R=0.791$; $P=1.57 \times 10^{-13}$) (Table 1).

We tested the age-predictive multiple linear regression model using simultaneously all the 5 CpG sites in the training set of 59 blood samples of living individuals. Although individually all the 5 CpG sites showed strong and significant associations with age (Table 1 & Supplementary Figs. S2A & S3A), in the multivariate analysis the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values (Supplementary Table S2), which could reveal signs of multicollinearity between variables. In concordance, when we applied the stepwise linear regression analysis, the same three significant CpG sites were chosen (at *ELOVL2*, *FHL2* and *C1orf132* genes). Thus, we built a final model with these 3 CpGs, which reveals a higher age correlation value ($R=0.982$), explaining 96.3% of age variation (adjusted $R^2=0.963$), highly significant ($P=7.315 \times 10^{-38}$) (Table 2). Applying this model, the age prediction for each individual was obtained through the formula: $(38.751)+61.058 \times \text{DNA methylation level } ELOVL2 + 80.021 \times \text{DNA methylation level } FHL2 - 47.631 \times \text{DNA methylation level } C1orf132$. The developed APM enabled us to estimate age with a correlation between predicted and chronological ages of 0.969 (Spearman correlation coefficient, $r=0.969$) (Fig. 1A) and a MAD from chronological age of 4.25 years.

The accuracy of the model was tested by a 4-fold cross validation. A set of samples (a 'fold') was randomly removed from the training set of 59 blood samples as validation set and four independent multiple linear regressions were performed on the remaining samples (training set). Subsequently, each prediction model is used to predict the age of the removed samples (validation set). The MAD between predicted and chronological ages were obtained in the validation set for each of the 4 independent multiple linear regressions and the calculated mean value reveal an average MAD value of 4.75 years amongst the four test sets, very close to the MAD of 4.25 years from the whole training set (Table 2). As an additional validation, we split the dataset into 2 similar sets of 30 and 29 samples each (training and validation set) and re-fitted the multivariate linear regression model on the training set. This allowed us to obtain an independent

Table 1

Simple linear regression statistics of the 5 CpGs at the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes in blood samples from living, deceased and in the overall (living and deceased) sample sets of Portuguese individuals.

Locus	Chromosomal location (GRCh38)	Group	N	R	R ²	P-value
<i>ELOVL2</i>	Chr6:11044628	Living	56	0.951	0.904	3.58×10^{-29}
		Deceased	62	0.791	0.626	2.04×10^{-14}
		Overall	118	0.919	0.845	7.60×10^{-49}
<i>FHL2</i>	Chr2:105399282	Living	59	0.946	0.895	1.49×10^{-29}
		Deceased	62	0.654	0.428	8.16×10^{-9}
		Overall	121	0.874	0.764	3.68×10^{-39}
<i>C1orf132</i>	Chr1:207823681	Living	59	-0.924	0.854	1.67×10^{-25}
		Deceased	60	-0.591	0.350	6.56×10^{-7}
		Overall	119	-0.834	0.695	6.15×10^{-32}
<i>TRIM59</i>	Chr3:160450189	Living	59	0.910	0.828	2.04×10^{-23}
		Deceased	61	0.769	0.591	4.78×10^{-13}
		Overall	120	0.830	0.688	1.16×10^{-31}
<i>KLF14</i>	Chr7:130734355	Living	58	0.791	0.625	1.57×10^{-13}
		Deceased	59	0.568	0.322	0.000003
		Overall	117	0.731	0.534	8.33×10^{-21}

Abbreviations: N, number of samples; R, correlation coefficient.

Table 2

Multiple linear regression statistics of CpGs at the *ELOVL2*, *FHL2*, *C1orf132*, *KLF14* and *TRIM59* genes in blood samples from living, deceased and in the overall (living and deceased) sample sets of Portuguese individuals.

Model	Group	R	R ² (adjusted R ²)	P-value	MAD	MAD 4-fold Cross
APM ^a	Living	0.982	0.965 (0.963)	7.32×10^{-38}	4.25	4.75
APM ^b	Deceased	0.899	0.808 (0.793)	1.07×10^{-18}	5.36	6.13
APM ^c	Overall	0.963	0.928 (0.925)	1.03×10^{-61}	4.97	5.25

Abbreviations: R, correlation coefficient; MAD, mean absolute deviation (years) between chronological and predicted ages; APM, age prediction model.

^a APM (*ELOVL2*, *FHL2* and *C1orf132* genes).

^b APM (*ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes).

^c APM (*ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes).

MAD value for the training set of 3.86 years. When this model was applied to the validation set, MAD of 4.93 years was obtained. Both independent MAD values were very close to the MAD of 4.25 obtained from the whole data set (59 individuals).

3.2. Comparing DNA methylation values between different populations

When we compare DNA methylation levels obtained from the multiplex methylation SNaPshot assay in blood samples from Portuguese and the Koreans [17] (Supplementary Table S3), all the CpG sites showed strong correlation with age ($0.763 < R < 0.951$), except for *C1orf132* ($R = -0.637$) in Koreans. The higher correlation value with age was obtained for *FHL2* in Koreans ($R = 0.893$) and for *ELOVL2* in Portuguese ($R = 0.951$). Moreover, in the multivariate analysis the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values ($P > 0.05$) in Portuguese individuals (data not shown), while in Koreans all the 5 CpG sites showed significant age correlation values ($P < 0.001$) [17]. Meanwhile, we tested the original model of Jung et al. [17] with our methylation data from 59 living individuals and a MAD value of 15.26 years was obtained.

3.3. DNA methylation analysis in blood samples from deceased individuals

A total of 62 blood samples from deceased individuals (aged 28–86 years-old; 49 males and 13 females) were tested using the same multiplex methylation SNaPshot assay. The CpG site in the *ELOVL2* showed the strongest age correlation ($R = 0.791$; $P = 2.04 \times 10^{-14}$), followed by *TRIM59* ($R = 0.769$; $P = 4.78 \times 10^{-13}$), *FHL2* ($R = 0.654$; $P = 8.16 \times 10^{-9}$), *C1orf132* ($R = -0.591$; $P = 6.56 \times 10^{-7}$) and finally the *KLF14* that, similar to living individuals, showed the lower age correlation value ($R = 0.568$; $P = 0.000003$) (Table 1).

We tested the age-prediction multiple linear regression model using simultaneously the 5 CpG sites, however the predictor CpG site at *KLF14* gene revealed a non-significant age correlation value (Supplementary Table S2). Furthermore, we used the stepwise regression to select the best model and the 4 CpG sites at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes were chosen. The final model constructed with these 4 predictor variables showed high age correlation coefficients ($R = 0.899$), explaining 79.3% of age variation (adjusted $R^2 = 0.793$), highly significant ($P = 1.07 \times 10^{-18}$) (Table 2). The developed formula obtained with the regression coefficients to calculate age was the following: $(14.914) + 63.627 \times \text{DNA methylation level } ELOVL2 + 40.299 \times \text{DNA methylation level } FHL2 + (-24.185) \times \text{DNA methylation level } C1orf132 + 57.717 \times \text{DNA methylation level } TRIM59$. The model showed a strong correlation between predicted and chronological ages (Spearman correlation coefficient, $r = 0.916$), with a MAD from chronological age of 5.36 years (Fig. 1B).

The 4-fold cross validation in the whole dataset of 62 blood samples from deceased individuals allowed to estimate an averaged MAD of 6.13 years, close to the MAD of 5.36 from the total training data set

(Table 2). The validation by splitting the sample in two equal sets of 31 samples each (training and validation set) allowed to obtain an independent MAD value for the training set of 6.15 years. The model was applied to the validation set and a MAD of 5.66 years was obtained. Both independent MAD values were very close to the MAD of 5.36 years obtained from the whole training data set (62 individuals).

3.4. DNA methylation analysis in blood samples from living and deceased individuals

When we combine living and deceased individuals obtaining a whole dataset of 121 blood samples (aged 1–94 years-old, 71 males, 50 females), considering all the 5 CpG sites for the multivariate analysis, a non-significant age correlation value was observed for *KLF14* (Supplementary Table S2). In concordance, the stepwise regression analysis also excluded *KLF14* from the model. Therefore, the APM selecting the 4 CpG sites at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes, showed the multiple regression age correlation coefficient $R = 0.963$, explaining 92.5% of age variation (adjusted $R^2 = 0.925$), highly significant ($P = 1.025 \times 10^{-61}$) (Table 2). The developed formula obtained with regression coefficients to calculate age was the following: $(17.936) + 66.925 \times \text{DNA methylation level } ELOVL2 + 52.009 \times \text{DNA methylation level } FHL2 + (-30.886) \times \text{DNA methylation level } C1orf132 + 44.391 \times \text{DNA methylation level } TRIM59$. The model showed a strong correlation between predicted and chronological ages (Spearman correlation coefficient, $r = 0.952$), with a MAD of 4.97 years (Fig. 1C).

A 4-fold cross validation using the whole dataset of 121 individuals allowed to estimate an averaged MAD of 5.25 years, close to the MAD of 4.97 from the whole data set (Table 2). The validation by splitting the sample in two sets of 61 and 60 samples each (training and validation set by using the previous groups) allowed to obtain an independent MAD value for the training set of 5.35 years. Applying the model to the validation set, a MAD of 5.35 years was obtained. Both independent MAD values were very close to the MAD of 4.97 years obtained from the whole data set (121 individuals).

3.5. Differences between predicted and chronological ages increase with age

The MAD values between predicted and chronological ages increase with the increasing of age in both blood sample sets of living and deceased individuals (Fig. 2). In living individuals, the MAD value was the largest for the age group > 61 years (MAD = 5.70 years) and the smallest for age group < 18 years (MAD = 3.03 years) (Table 3). In deceased individuals, the largest MAD was obtained in individuals of the age group 72–88 years (MAD = 6.24 years), while the smallest MAD was observed in the age group 28–51 years (MAD = 5.00 years) (Table 3).

4. Discussion

A number of DNA methylation markers have been proposed as predictors of chronological age. However, for better application of

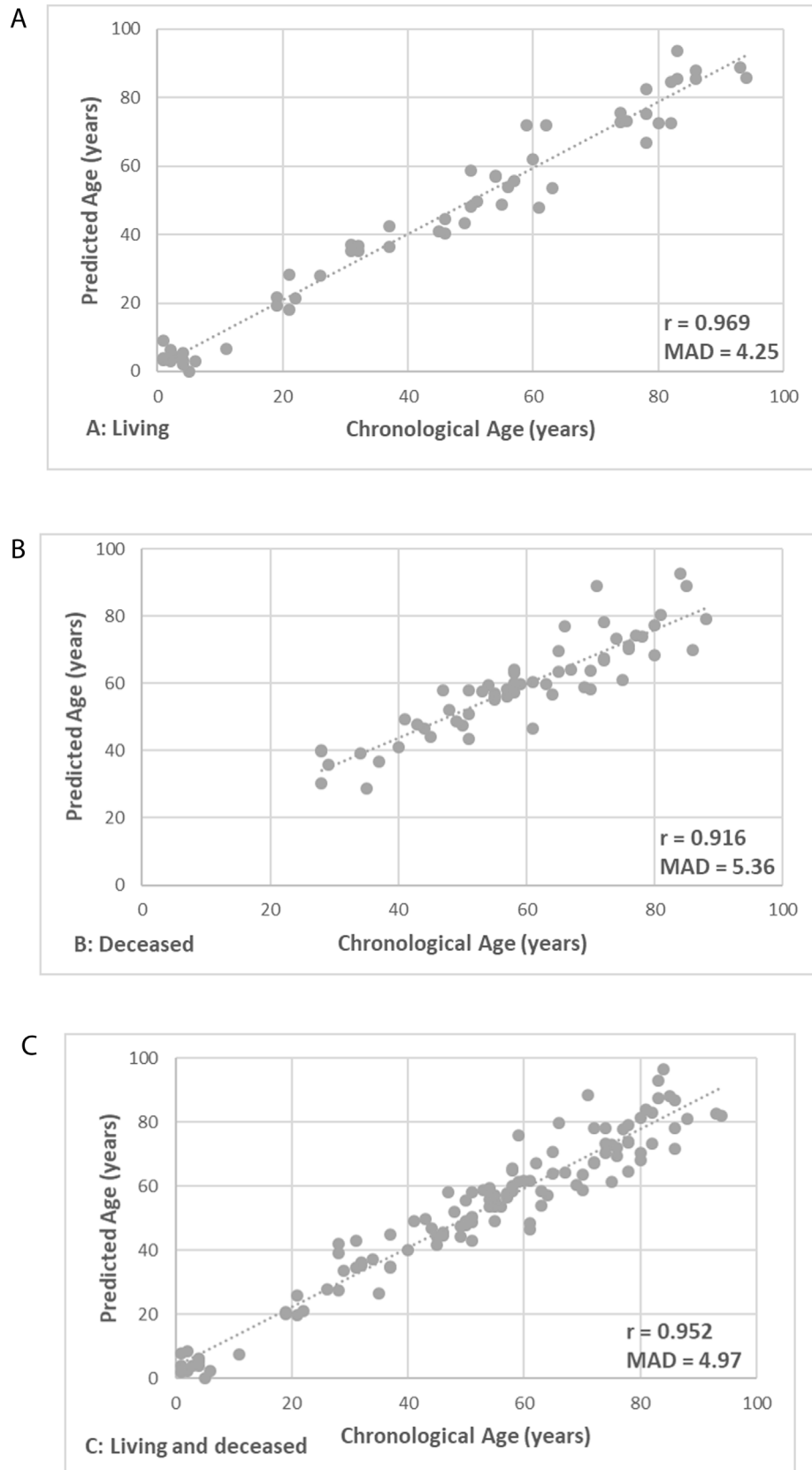


Fig. 1. Predicted versus chronological ages in the sample sets of blood samples. A) APM developed with the 3 CpGs at *ELOVL2*, *FHL2* and *C1orf132* genes in living individuals. B) APM developed with 4 CpGs at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in deceased individuals. C) APM developed with the 4 CpGs at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in the overall sample of living and deceased individuals. MAD and r (Spearman correlation coefficient) values are plotted in each graph.

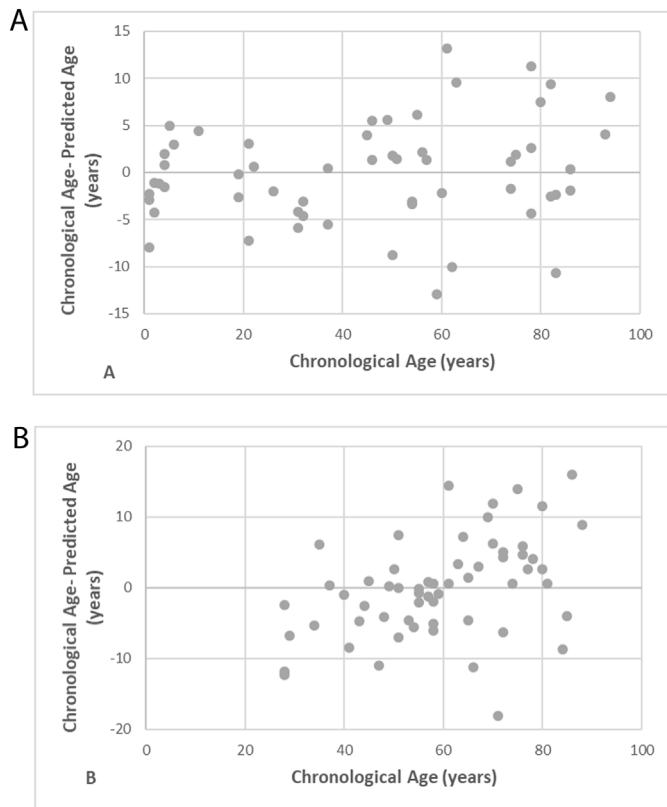


Fig. 2. Differences between chronological and predicted ages (years) plotted against chronological age (years) in blood samples from living individuals (A) and from deceased individuals (B).

Table 3

Evaluation of mean absolute deviation (MAD) between chronological and predicted ages in several age range groups in living and deceased individuals.

Group	Age range	N	MAD
Living individuals			
G1	<18 years	11	3.03
G2	19–39 years	12	3.28
G3	40–60 years	14	4.26
G4	>61 years	18	5.70
Deceased individuals			
G1	28–51 years	19	5.00
G2	52–71 years	24	5.07
G3	72–88 years	16	6.24

Abbreviations: N, number of samples.

developed approaches to forensic casework sample analysis it is important to further test the proposed markers and bisulfite sequencing methods. Therefore, in the present study we tested the experiments proposed by Jung et al. [17] using a SNaPshot assay in an independent set of blood samples of healthy Portuguese individuals using the same markers and methodology. Moreover, we evaluated DNA methylation levels through the SNaPshot method in a set of blood samples from deceased individuals.

In the current study, strong age correlation values were obtained for the 5 CpG sites at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes in blood samples from Portuguese living individuals. The final APM proposed by Jung et al. [17] included all the 5 genes *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14* and *FHL2*, demonstrating high age prediction accuracy in the Korean population (MAD = 3.174). In the Portuguese study sample, our proposed final

APM was built only with those 3 CpGs, at *ELOVL2*, *FHL2* and *C1orf132* genes, that showed statistical significance in the multivariate regression analysis. Nevertheless, this final APM exhibited an age prediction accuracy (MAD = 4.25) similar to the model of Jung et al. [17]. Thus, our study replicated the age association test of Jung et al. [17] in an independent validation set, evidencing that the performance of these markers through the SNaPshot methodology is consistent across different populations.

However, comparing our data with that obtained in the Korean population some differences can be found in the extent of the age association for the targeted loci. In particular, CpG at *ELOVL2* showed the strongest age correlation in Portuguese individuals, while in Koreans the CpG at *FHL2* showed the best result [17]. On the other hand, the lowest correlation value was obtained for the *KLF14* CpG in Portuguese but for the *C1orf132* CpG in Koreans. Moreover, when we tested the original algorithm model of Jung et al. [17] with our obtained methylation data from living individuals, a lower accurate MAD value of 15.26 years was obtained. Similarly, a study by Cho et al. [8] in the Korean population replicating a study of Zbieć-Piekarska et al. [5] in the Polish population using pyrosequencing found age correlation differences in specific markers. Such kind of differences suggest the possibility that some effects in methylation levels may be population specific. This phenomenon supports the notion that specific markers can be more adequate to different population groups to explain age-related DNA methylation variance. This points to the usefulness of replication and validation studies of proposed markers and genotyping methods in different populations and datasets before forensic applications.

In our study, the final model constructed with 4 predictor variables at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes for samples of deceased individuals, also exhibited age prediction accuracy with a MAD from chronological age of 5.36 years. To the best of our knowledge, only three studies have focused in blood samples from deceased individuals for development of age prediction models based on DNA methylation [4,12,18]. These three studies evaluated DNA methylation levels among a number of different loci including *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* genes. When comparing blood samples from living and deceased individuals, two of these studies [4,18] suggested similar distributions of DNA methylation patterns. However, in our study, the individual analysis of DNA methylation levels showed higher age correlation values in living individuals ($0.79 \leq R \leq 0.95$) compared to deceased individuals ($0.57 \leq R \leq 0.79$). Moreover, in the multivariate analysis the CpG sites at *KLF14* and *TRIM59* genes showed non-significant age correlation values in living individuals, while in deceased individuals only the *KLF14* CpG site showed a non-significant p-value (Supplementary Table S2). Thus, this data suggest that *postmortem* changes can occur in the methylation levels. Nevertheless, a final model for age prediction addressing the whole sample set of living and deceased individuals constructed with the same 4 CpGs in *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes, exhibited high age prediction accuracy explaining 92.5% of variation in age, with a MAD of 4.97 years.

Our study showed that the prediction accuracy depends on the chronological age of individuals, both in living and deceased. The higher MAD values were obtained in older ages, in concordance with previous studies [4–6,18]. This mean that individual differences in the rate of methylation change occurs with age, being slight in youths and accumulating with age, enabling that the age prediction models are more accurate in younger than in older [4,18]. Moreover, our reports showed non-significant gender DNA methylation differences, similar to previous studies [4,5].

In conclusion, we validated and replicated in blood samples of Portuguese individuals the SNaPshot assay of Jung et al. [17] by evaluating DNA methylation levels in 5 specific CpG sites from the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes. Although using a

different APM built with only 3 CpGs at *ELOVL2*, *FHL2* and *C1orf132* genes, similar high prediction accuracies were obtained (MAD=4.25 in Portuguese and MAD=3.174 in Koreans). Some differences in the methylation patterns were observed between the two populations (Koreans and Portuguese) suggesting the possibility that some effects in DNA methylation levels of age markers might be population specific. The possibility that *postmortem* changes could alter the methylation status among specific loci was also suggested by our data. Nevertheless, the development of independent age prediction models with 4 CpGs at *ELOVL2*, *FHL2*, *C1orf132* and *TRIM59* genes in blood samples from deceased individuals, as well as in the whole data set of living and deceased individuals, showed good levels of age prediction accuracy (5.36 and 4.97 years of MAD, respectively). Our study showed the usefulness of the multiplex methylation SNaPshot assay for forensic analysis in blood samples of living and deceased individuals. This methodology seems promising in forensic fields because of its capacity for multiplexing analysis, investigating simultaneously the DNA methylation level across several specific CpGs.

CRedit authorship contribution statement

Helena Correia Dias: Data curation, Software, Writing - original draft, Methodology. **Cristina Cordeiro:** Writing - review & editing, Validation, Methodology. **Janet Pereira:** Methodology. **Catarina Pinto:** Methodology. **Francisco Corte Real:** Writing - review & editing, Supervision, Validation. **Eugénia Cunha:** Writing - review & editing, Supervision, Validation. **Licínio Manco:** Writing - review & editing, Conceptualization, Methodology, Supervision.

Declarations of interest

None.

Acknowledgements

This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) (FCT-PEst-OE/SADG/UI0283/2019). H.D. has a PhD grant from FCT (SFRH/BD/117022/2016). The authors thank all the supporting for the present studying included the people who collected blood samples from deceased individuals in INMLCF and also the people that collected blood samples from Bodies Donated to Science in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto*. We thank Dr. Veronica Gomes from IPATIMUP for technical assistance in the SNaPshot assay.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at <https://doi.org/10.1016/j.forsciint.2020.110267>.

References

- [1] A. Freire-Aradas, C. Phillips, M.V. Lareu, Forensic individual age estimation with DNA: from initial approaches to methylation tests, *Forensic Sci. Rev.* 29 (2017) 121–144.
- [2] N. Goel, P. Karira, V.K. Garg, Role of DNA methylation in human age prediction, *Mech. Ageing Dev.* 166 (2017) 33–41.
- [3] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van de Voorde, R. Decorte, A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples, *Forensic Sci. Int.: Genet. Suppl. Ser.* 5 (2015) e144–e145.
- [4] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van de Voorde, R. Decorte, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, *Epigenetics* 10 (2015) 922–930.
- [5] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Z. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki, Development of a forensically useful age prediction method based on DNA methylation analysis, *Forensic Sci. Int.: Genet.* 17 (2015) 173–179.
- [6] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, Z. Makowska, A. Spas, A. Parys-Proszek, K. Kucharczyk, R. Płoski, W. Branicki, Examination of DNA methylation status of the *ELOVL2* marker may be useful for human age prediction in forensic science, *Forensic Sci. Int.: Genet.* 14 (2015) 161–167.
- [7] M. Eipel, F. Mayer, T. Arent, M.R.P. Ferreira, C. Birkhofer, I.G. Costa, et al., Epigenetic age predictions based on buccal swabs are more precise in combination with cell type - specific DNA methylation signatures, *Aging* 8 (2016) 1034–1048.
- [8] S. Cho, S.-E. Jung, S.R. Hong, E.H. Lee, J.H. Lee, S.D. Lee, et al., Independent validation of DNA-based approaches for age prediction in blood, *Forensic Sci. Int.: Genet.* 29 (2017) 250–256.
- [9] H. Correia Dias, C. Cordeiro, F. Corte Real, E. Cunha, L. Manco, Age estimation based on DNA methylation using blood samples from deceased individuals, *J. Forensic Sci.* (2019) 1–6.
- [10] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares de Cal, J. Álvarez-Dios, J. Ansedo-Bermejo, M. Torres-Español, P.M. Schneider, E. Pošpiech, W. Branicki, Á. Carracedo, M.V. Lareu, Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system, *Forensic Sci. Int.: Genet.* 24 (2016) 65–74.
- [11] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der, P. Zwalm Henneman, et al., Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression, *Forensic Sci. Int.: Genet.* 31 (2017) 19–28.
- [12] J. Naue, T. Sängler, H.C.J. Hoefsloot, S. Lutz-Bonengel, A.D. Kloosterman, P.J. Verschure, Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing, *Forensic Sci. Int.: Genet.* 36 (2018) 152–159.
- [13] A. Aliferi, D. Ballard, M.D. Gallidabino, H. Thurtle, L. Barron, D. Syndercombe Court, DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models, *Forensic Sci. Int.: Genet.* 37 (2018) 215–226.
- [14] H.Y. Lee, S.-E. Jung, Y.N. Oh, A. Choi, W.I. Yang, K.-J. Shin, Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study, *Forensic Sci. Int.: Genet.* 19 (2015) 28–34.
- [15] S.R. Hong, S.E. Jung, E.H. Lee, K.J. Shin, W.I. Yang, H.Y. Lee, DNA methylation based age prediction from saliva: high age predictability by combination of 7 CpG markers, *Forensic Sci. Int.: Genet.* 29 (2017) 118–125.
- [16] S.R. Hong, K.-J. Shin, S.-E. Jung, E.H. Lee, H.Y. Lee, Platform-independent models for age prediction using DNA methylation data, *Forensic Sci. Int.: Genet.* 38 (2019) 39–47.
- [17] S.E. Jung, S.M. Lim, S.R. Hong, E.H. Lee, K.J. Shin, H.Y. Lee, DNA methylation of the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*/*MIR29B2C*, and *TRIM59* genes for age prediction from blood, saliva, and buccal swab samples, *Forensic Sci. Int.: Genet.* 38 (2019) 1–8.
- [18] Y. Hamano, S. Manabe, C. Morimoto, S. Fujimoto, M. Ozeki, K. Tamaki, Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting, *Leg. Med.* 21 (2016) 5–10.
- [19] R. Eisma, C. Lamb, R.W. Soames, From Formalin to thiel embalming: what changes? One anatomy department's experiences, *Clin. Anat.* 26 (2013) 564–571.






DNA methylation age estimation from human bone and teeth

Helena Correia Dias , Francisco Corte-Real , Eugénia Cunha & Licínio Manco


To cite this article: Helena Correia Dias , Francisco Corte-Real , Eugénia Cunha & Licínio Manco (2020): DNA methylation age estimation from human bone and teeth, Australian Journal of Forensic Sciences, DOI: [10.1080/00450618.2020.1805011](https://doi.org/10.1080/00450618.2020.1805011)

To link to this article: <https://doi.org/10.1080/00450618.2020.1805011>

 [View supplementary material](#) 

 Published online: 17 Aug 2020.

 [Submit your article to this journal](#) 

 Article views: 44

 [View related articles](#) 

 [View Crossmark data](#) 



DNA methylation age estimation from human bone and teeth

Helena Correia Dias^{a,b,c}, Francisco Corte-Real^{c,d}, Eugénia Cunha^{b,c}
and Licínio Manco^{id}^a

^aResearch Centre for Anthropology and Health (CIAS), Department of Life Sciences, University of Coimbra, Coimbra, Portugal; ^bCentre for Functional Ecology (CEF), Laboratory of Forensic Anthropology, Department of Life Sciences, University of Coimbra, Coimbra, Portugal; ^cNational Institute of Legal Medicine and Forensic Sciences, Coimbra, Portugal; ^dFaculty of Medicine, University of Coimbra, Coimbra, Portugal

ABSTRACT

Bones and teeth are valuable sources of information in forensic contexts. However, few studies have considered these tissues for DNA methylation (DNAm) analyses in age estimation. The main goal of the present study was to develop DNAm-based age prediction models (APMs) specific for bones and teeth. DNA samples from Portuguese individuals were evaluated through bisulphite Sanger sequencing and a multiplex SNaPshot assay to analyse relationships between DNAm and age. In bones our results allowed to build a final APM through Sanger sequencing with six CpGs at genes *ELOVL2*, *EDARADD* and *MIR29B2C*, explaining 92.5% of age variation, with a mean absolute deviation (MAD) from chronological age of 2.56 years. For teeth, we were unable to build a final multi-locus APM; the best *FHL2* CpG4 age predictor explained 41.3% of age variance with a MAD of 11.35 years. The SNaPshot assay allowed to build two final dual-locus APMs with CpGs at *FHL2* and *KLF14* for bones, explaining 57.6% of age variation (MAD of 7.18 years), and with CpGs at *ELOVL2* and *KLF14* for teeth, explaining 76.4% of age variation (MAD of 7.07 years). Our study showed the usefulness of bone and tooth samples in evaluation of DNAm levels for age estimation purposes.

ARTICLE HISTORY

Received 7 May 2020
Accepted 28 June 2020

KEYWORDS

DNA methylation; age estimation; bisulphite Sanger sequencing; multiplex methylation SNaPshot; human teeth and bones

Introduction

Forensic anthropologists have to deal with human identification, not only in routine practice but also in exception scenarios, associated with mass disaster or crimes against humanity including war crimes and human rights violation. Age estimation is one of the most relevant questions in forensic contexts, being necessary both for living and deceased individuals.¹ The multidisciplinary approach to age estimation is paramount. Macroscopic and imagiological analyses in bones and teeth, followed by an appropriate mathematical approach are now standard procedures. But genetics and chemistry are also playing important roles in that respect. Yet, despite the existence of many biomolecular and chemical methods to predict age, there is no standard method that can be applied for all the forensic scenarios.

Nowadays, analysis of DNA methylation (DNAm) has currently arisen as a promising tool for predicting age.^{2,3} Many age-dependent DNAm markers have been identified in various tissues using several methodologies for target bisulphite sequencing including pyrosequencing,⁴⁻⁹ Sanger sequencing,¹⁰ EpiTyper,¹¹ massive parallel sequencing¹²⁻¹⁴ or SNaPshot.¹⁵⁻¹⁹ Although many CpG markers have been identified showing high correlations with chronological age, validation of methodologies and markers should be made for development of accurate age prediction models (APMs) potential useful in forensic analysis. Also, due to the highly tissue-specific nature of DNAm,^{20,21} the development of specific models for particular types of tissues should be tested.

To date, few studies have considered bones^{13,22} and teeth^{5,23,24} for DNAm analyses. Using massive parallel sequencing, Naue et al.¹³ obtained in bone samples from 29 individuals a moderate correlation between DNAm and age for CpGs at *ELOVL2*, *KLF14* and *TRIM59* genes (0.58, 0.51 and 0.61, respectively). More recently, Gopalan et al.²² using genome-wide DNAm data from 165 bone samples developed a powerful '37 bone clock CpGs' for age prediction based on CpG sites from *TRIM59*, *ELOVL2* and *KLF14* genes, among other. Giuliani et al.²³ investigated methylation data at *ELOVL2*, *FHL2* and *PENK* loci by MALDI-ToF mass spectrometry in 21 modern teeth and proposed an APM for cementum, dentin and pulp with the mean absolute deviation (MAD) between estimated and chronological ages of 2.45, 7.07 and 2.25 years, respectively. In concordance, Bekaert et al.⁵ evaluating DNAm levels by pyrosequencing in 29 dentin samples from living individuals, reported a multiple quadratic regression model with seven CpGs located at *PDE4C*, *ELOVL2* and *EDARADD* genes explaining 74% of age variance with a MAD of 4.86 years. A recent study by Márquez-Ruiz et al.²⁴ testing by bisulphite pyrosequencing the methylation levels in 65 tooth samples from individuals, obtained a significant positive age association for CpG sites at *ELOVL2* and *PDE4C* and developed an APM with nine CpGs from these two loci with a mean absolute error (MAE) of 5.08 years.

Considering these promising results, the main goal of the present study was to investigate DNAm information for age prediction purposes in bone and tooth samples from identified Portuguese individuals using the bisulphite polymerase chain reaction (PCR) Sanger sequencing and multiplex SNaPshot methodologies. Both methods used herein are semi-quantitative but have shown to be efficient and economical alternative tools for rapid quantification of DNAm in two previous studies from our group.^{10,19}

Material and methods

Sample collection

A set of 31 bone samples (aged 26–81 years old; 26 males, 5 females) were collected from identified deceased individuals during autopsies in *Serviço de Patologia Forense da Delegação do Centro e Sul*, after consulting RENNDA (*Registo Nacional de Não Dadores*). A set of 31 tooth samples (aged 26–94 years old; 10 males, 21 females) were collected from living individuals ($n = 23$) in dentist offices, after informed consent, and from Bodies Donated to Science (BDS) ($n = 8$) in *Departamento de Anatomia da Faculdade de Medicina da Universidade do Porto* before the embalming method of the body. The collection of teeth before the embalming ensures the control of any possible influence related to the process of conservation.

The study protocol was approved by the *Instituto Nacional de Medicina Legal e Ciências Forenses* (INMLCF) and by the Ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (n° 038-CE-2017).

DNA extraction, bisulphite conversion

Processing of bone and tooth samples was made in INMLCF, according to standard guidelines. Cleaning and grinding bone/teeth took place in a room designated exclusively for processing of human remains. All sample manipulations were performed in laminar flow chambers equipped with filters and UV lights. The procedure consists in the following steps: washing the sample type (bone and teeth) with bleach for 5 minutes, and then it was put in distilled water (ddH₂O) for another 5 minutes to remove residual bleach. A drill was used to remove other exogenous contaminants. After, we make some cuts (around 0.5 × 0.5 cm) in the fragment and it was put in a vial for grinding bone using a 6770 freezer/mill and nitrogen liquid. DNA extraction was performed using a robot with *PrepFiler Express BTA™ Forensic DNA Extraction Kit* (Applied Biosystems, Foster City, CA). DNA Quantification was made with the real-time polymerase chain reaction (PCR)-based kit *Quantifiler™ Human DNA Quantification Kit* (Applied Biosystems).

After extraction and quantification, genomic DNA was subjected to bisulphite conversion using *EZ DNA Methylation Gold Kit* (Zymo Research, Irvine, USA) according to the instructions of manufacturer. Briefly, 20 µl of genomic DNA (in a total amount of 200 to 400 ng) was treated with sodium bisulphite and modified DNA was extracted to a final volume of 10 µL.

Sanger sequencing

After bisulphite conversion, the modified DNA samples were submitted to PCR for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *MIR29B2C* using the *Qiagen Multiplex PCR kit* (Qiagen, Hilden, Germany) and sequenced with *Big-Dye Terminator v1.1 Cycle Sequencing kit* (Applied Biosystems), using primers and conditions previously described.¹⁰

SNaPshot assay

After bisulphite conversion, the modified DNA samples were submitted to a multiplex SNaPshot assay for 5 CpG sites at genes *ELOVL2*, *FHL2*, *KLF14*, *MIR29B2C* and *TRIM59* with the primers and conditions previously described in Jung et al.¹⁸ Particular conditions for multiplex PCR amplification and multiplex SBE (single-base extension) reactions were as previously described.¹⁹

Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Linear regression models were used to analyse relationships between methylation levels and chronological age. For Sanger sequencing, using the simple linear regression coefficients from the highest age-

correlated CpG sites from each gene we predicted age of individuals in bones and teeth. All the significant CpGs were combined for analysis using the stepwise regression approach for selection of the relevant variables to be included in final APMs. For the SNaPshot method, simple linear regression coefficients of each significant age-correlated CpG were used to predict age of individuals in bones and teeth. Methylation information of the significant age-correlated CpGs was used in the stepwise regression approach to select the predictor variables to be used in the final APMs. Spearman correlation coefficients and MAD values between predicted and chronological ages were calculated for training sets of bone and tooth samples. Validation of the final APMs was performed by threefold cross-validation that consists in removing randomly a set of samples from the training set and to develop three independent multiple linear regressions on the remaining samples. Subsequently, each APM is used to predict the age of the removed samples assigned as validation sets. An additional validation was performed by splitting the complete data set into two subsets (training and validation sets) and an independent regression was calculated for the training set and applied to the validation set.

Results

In the present work we report on methylation levels of 43 CpG sites located at *ELOVL2*, *FHL2*, *PDE4C*, *EDARADD* and *MIR29B2C* genes through the bisulphite Sanger sequencing methodology obtained in 29 fresh bone samples (4 females and 25 males; aged 26–80 years old) and in 31 tooth samples (23 from living and 8 from deceased individuals; 10 males and 21 females; aged 26–94 years old). Moreover, using the multiplex SNaPshot assay reported by Jung et al.¹⁸, methylation data from 5 CpG sites at *ELOVL2*, *FHL2*, *KLF14*, *MIR29B2C* and *TRIM59* genes were obtained from 31 fresh bone samples (26 males and 5 females; aged 26–81 years old) and from 24 tooth samples (16 from living and 8 from deceased individuals; 8 males and 16 females; aged 27–88 years old).

DNA methylation data obtained using bisulphite Sanger sequencing

The DNAm levels from bone samples analysed through bisulphite Sanger sequencing revealed positive correlations for all the CpG sites located at genes *ELOVL2*, *FHL2* and *PDE4C*, and negative correlations for the CpG sites on the remaining two genes *EDARADD* and *MIR29B2C* (Supplemental Table S1). Simple linear regression showed strong and significant age-correlation values ($0.70 < R < 0.90$) for almost all CpGs located at *ELOVL2* and for the three first CpG sites at *MIR29B2C* (Supplemental Table S1). The remaining markers showed negligible, weak or moderate age correlations. The strongest age-correlated value was observed for the *ELOVL2* CpG6 ($R = 0.852$; $P = 4.64 \times 10^{-9}$), following by *MIR29B2C* CpG1 ($R = -0.834$; $P = 1.93 \times 10^{-8}$) (Table 1; Supplemental Figure S1). For the remaining genes, the best markers showed moderate age correlation values: *FHL2* CpG1 ($R = 0.692$, $P = 0.000032$), *PDE4C* CpG2 ($R = 0.690$; $P = 0.000049$) and *EDARADD* CpG3 ($R = -0.561$; $P = 0.001564$) (Table 1; Supplemental Figure S1). Simple linear regression APMs based on these individual CpGs revealed MAD values from chronological age of 5.73 years for *ELOVL2* CpG6, 6.35 years for *MIR29B2C* CpG1, 8.39 years for *FHL2* CpG1, 8.07 years for *PDE4C* CpG2 and 9.23 years for *EDARADD* CpG3 (Table 1). We selected the 29 CpG sites which showed individual significant association with age, located at *ELOVL2* (9 CpGs), *FHL2* (7 CpGs), *EDARADD* (3 CpGs), *PDE4C* (5 CpGs) and *MIR29B2C* (5 CpGs), and we

Table 1. Simple and multiple linear regression statistics of the five best CpGs located at *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *MIR29B2C* genes using methylation information from bone samples obtained from bisulphite Sanger sequencing.

Locus	CpG	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
Simple linear regression									
<i>ELOVL2</i>	CpG6	Chr6:11044644	29	0.852	0.725	0.715	7.24	4.64×10^{-9}	5.73
<i>MIR29B2C</i>	CpG1	Chr1:207823681	29	-0.834	0.695	0.684	7.63	1.93×10^{-8}	6.35
<i>FHL2</i>	CpG1	Chr2:105399282	29	0.692	0.479	0.460	9.97	0.000032	8.39
<i>PDE4C</i>	CpG2	Chr19:18233133	28	0.690	0.476	0.456	10.19	0.000049	8.07
<i>EDARADD</i>	CpG3	Chr1:236394382	29	-0.561	0.314	0.289	11.45	0.001564	9.23
Multiple linear regression									
APM (<i>ELOVL2</i> CpG5 + <i>ELOVL2</i> CpG6 + <i>ELOVL2</i> CpG7 + <i>MIR29B2C</i> CpG1 + <i>EDARADD</i> CpG3 + <i>EDARADD</i> CpG4)			29	0.970	0.941	0.925	3.71	2.097×10^{-12}	2.56

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

applied the stepwise linear regression analysis to select the relevant variables to be used in a final APM (Supplemental Table S2). The selected six CpGs located at *ELOVL2* (CpG5, CpG6, CpG7), *EDARADD* (CpG3, CpG4) and *MIR29B2C* (CpG1) revealed in the multiple regression analysis a high age correlation coefficient ($R = 0.970$), explaining 92.5% of age variance, highly significant ($P = 2.097 \times 10^{-12}$) (Table 1). Age prediction for each individual bone sample was estimated according to the equation developed with the regression coefficients as present in Supplemental Table S2: $129.912 - 66.051 \times \text{DNAm levels } \textit{MIR29B2C} \text{ CpG1} - 136.346 \times \text{DNAm levels } \textit{EDARADD} \text{ CpG3} + 62.928 \times \text{DNAm levels } \textit{EDARADD} \text{ CpG4} + 67.573 \times \text{DNAm levels } \textit{ELOVL2} \text{ CpG5} + 144.915 \times \text{DNAm levels } \textit{ELOVL2} \text{ CpG6} - 137.429 \times \text{DNAm levels } \textit{ELOVL2} \text{ CpG7}$. The developed APM enabled us to estimate age with a correlation between predicted and chronological ages of 0.957 (Figure 1) and a MAD from chronological age of 2.56 years (Table 1). The accuracy of the model was tested by a threefold cross-validation using the 29 bone samples. In this validation, a 'fold' (set of samples with 10 or 9 samples) was randomly removed from the training set as a validation set and three independent multiple linear regressions models were performed on the remaining samples (training sets). Each independent multiple linear regression predictive model was tested in the removed samples (validation sets) allowing to obtain an averaged MAD from chronological age of 3.77 years, a value very close to the MAD of 2.56 of the whole training set. Additionally, a second validation method by splitting the sample in two sets of 16 and 13 samples (training and validation sets) revealed an independent MAD value for the training set of 2.03 years. The multiple linear regression equation was applied to the validation set allowing to obtain a MAD of 3.15 years.

The DNAm levels from tooth samples evaluated through the Sanger sequencing methodology showed no significant correlation with age for any CpG site at *EDARADD* and *MIR29B2C* genes (Supplemental Table S3). The CpG sites at the remaining genes *ELOVL2*, *FHL2*, *PDE4C* showed positive but weak correlations between DNAm and age: the strongest site was *FHL2* CpG4 ($R = 0.658$, $P = 0.000078$), followed by *PDE4C* CpG1 ($R = 0.474$, $P = 0.013$) and *ELOVL2* CpG3 ($R = 0.379$, $P = 0.036$) (Table 2; Supplemental Figure S2). We tested the age prediction multiple linear regression model using simultaneously these three CpG sites; however, only CpG4 site at *FHL2* gene revealed a significant age correlation value ($P = 0.002$). Furthermore, we used the stepwise regression approach selecting the 12 significant age-correlated CpG sites located at *ELOVL2* (2 CpGs), *FHL2* (6 CpGs) and *PDE4C* (4 CpGs) but only the same *FHL2* CpG4 was chosen. Age prediction for

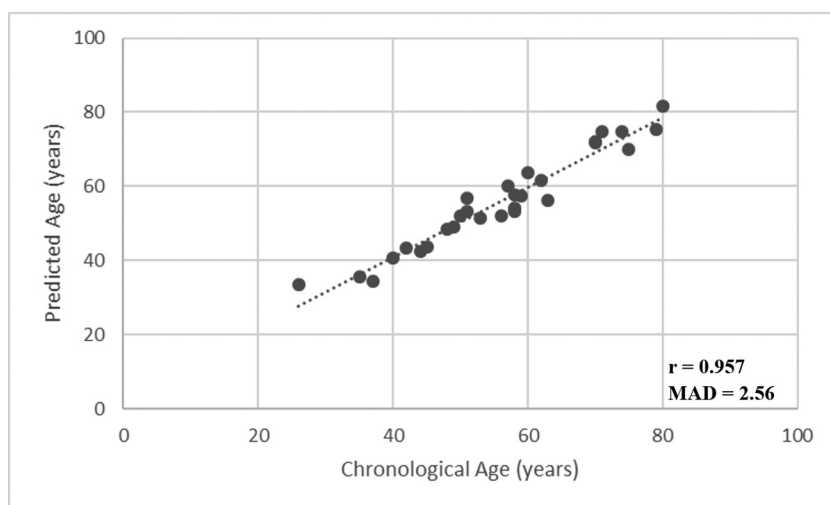


Figure 1. Plot of predicted vs. chronological ages using the final APM based on 6 CpGs located at *ELOVL2* (CpG5, CpG6, CpG7), *EDARADD* (CpG3, CpG4) and *MIR29B2C* (CpG1) built with methylation information from 29 fresh bone samples obtained with bisulphite PCR Sanger sequencing. Spearman correlation coefficient and MAD value from chronological age were plotted in the graph.

Table 2. Simple linear regression analysis of the three best CpG sites located at *FHL2*, *PDE4C* and *ELOVL2* genes using methylation information from tooth samples from living and deceased individuals obtained from bisulphite Sanger sequencing.

Locus	CpG	Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
Simple linear regression									
<i>FHL2</i>	CpG4	Chr2:105399297	30	0.658	0.433	0.413	14.57	0.000078	11.35
<i>PDE4C</i>	CpG1	Chr19:18233139	27	0.474	0.224	0.193	17.54	0.012588	14.58
<i>ELOVL2</i>	CpG3	Chr6:11044634	31	0.379	0.143	0.114	18.32	0.035738	15.22

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

each individual sample estimated according to the individual regression coefficients obtained from the 30 tooth samples with positive PCR amplification for the highest age associated marker *FHL2* CpG4 was as follows: $(-114.989) + 239.863 \times \text{DNAm levels } FHL2 \text{ CpG4}$ (Supplemental Table S4). The predicted age captured from the 30 tooth samples allowed to obtain a moderate correlation between predicted and chronological ages of 0.589 (Figure 2), with a MAD from the chronological age of 11.35 years (Table 2).

DNA methylation data obtained using the multiplex SNaPshot methodology

Five specific CpGs located at genes *ELOVL2*, *FHL2*, *KLF14*, *MIR29B2C* and *TRIM59* were simultaneously measured through a multiplex SNaPshot assay in 31 fresh bone samples from deceased individuals. Positive correlations with age were observed for CpGs at genes *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* and a negative correlation was obtained for *MIR29B2C* locus (Supplemental Figure S3). Among the five markers, the CpG site in the *FHL2* locus showed the strongest age correlation ($R = 0.708$, $P = 0.000008$), followed by *TRIM59* ($R = 0.633$, $P = 0.000129$), *ELOVL2* ($R = 0.619$, $P = 0.000202$), *KLF14* ($R = 0.540$, $P = 0.001708$) and *MIR29B2C*

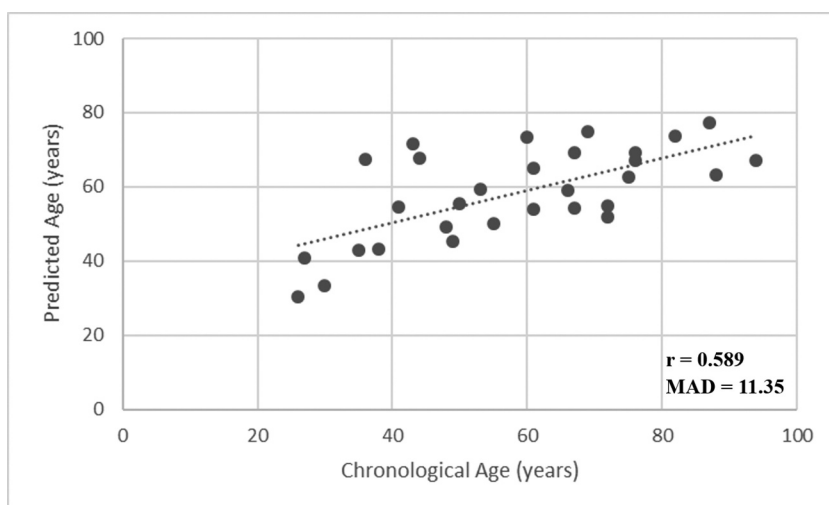


Figure 2. Plot of predicted vs. chronological ages using the APM based on the high age associated marker *FHL2* CpG4 built with methylation information from 30 tooth samples obtained with bisulphite PCR Sanger sequencing. Spearman correlation coefficient and MAD value from chronological age were plotted in the graph.

($R = -0.507$, $P = 0.003640$) (Table 3; Supplemental Figure S3). Single APMs using each CpG site revealed MAD values from the chronological age of 7.95 years for *FHL2*, 8.50 years for *ELOVL2*, 8.86 years for *TRIM59*, 9.29 years for *KLF14* and 9.70 years for *MIR29B2C* genes (Table 3). Fitting the stepwise regression analysis by using simultaneously the 5 CpG sites, only CpG sites at *FHL2* and *KLF14* genes were chosen. The final multiple dual-locus model built with CpG sites at *FHL2* and *KLF14* showed a moderate age correlation value ($R = 0.777$; $P = 0.000002$), explaining 57.6% of age variation (Table 3). The developed formula obtained with the regression coefficients to predict age was as follows: $15.727 + 105.392 \times DNAm\ levels\ FHL2 + 154.672 \times DNAm\ levels\ KLF14$ (Supplemental Table S5). The two-locus model showed a moderate correlation of 0.746 between predicted and chronological ages (Figure 3), with a MAD from chronological age of 7.18 years (Table 3). The threefold cross-validation to test the accuracy of the model, showed an averaged MAD from the chronological age for the three independent validation sets of 7.84 years, a value very close to the MAD of 7.18 of the whole training set. The validation by splitting the sample in two sets of 16 and 15 samples each (training and validation sets) allowed to obtain an independent MAD from the chronological age for the training set of 5.40 years and for the validation set of 9.35 years.

The tooth samples from living and deceased individuals evaluated using the same multiplex methylation SNaPshot assay showed positive and significant age correlations for *ELOVL2*, *KLF14* and *TRIM59* genes (Table 4; Supplemental Figure S4). The CpG sites in *FHL2* and *MIR29B2C* genes showed a lower or no significant age correlation (Table 4) and were excluded for further analysis. DNAm levels in *KLF14* locus showed the strongest age-correlation ($R = 0.728$, $P = 0.000084$), following by *ELOVL2* ($R = 0.685$, $P = 0.000311$) and *TRIM59* ($R = 0.665$, $P = 0.000389$) (Table 4). Using these sites in simple linear regression models we calculated values of MAD from chronological age of 9.68 years for *KLF14*, 11.27 years for *ELOVL2* and 11.51 years for *TRIM59*. The stepwise regression approach allowed to select a final

Table 3. Simple and multiple linear regression statistics at the 5 CpGs located at *ELOVL2*, *FHL2*, *MIR29B2C*, *KLF14* and *TRIM59* loci using methylation information from bone samples obtained from a multiplex methylation SNaPshot analysis.

Locus	CpG Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
Simple linear regression								
<i>FHL2</i>	Chr2:105399282	31	0.708	0.501	0.484	10.26	0.000008	7.95
<i>TRIM59</i>	Chr3:160450189	31	0.633	0.402	0.381	11.23	0.000129	8.86
<i>ELOVL2</i>	Chr6:11044628	31	0.619	0.384	0.363	11.40	0.000202	8.50
<i>KLF14</i>	Chr7:130734355	31	0.540	0.292	0.267	12.23	0.001708	9.29
<i>MIR29B2C</i>	Chr1:207823681	31	-0.507	0.257	0.231	12.53	0.003640	9.70
Multiple linear regression								
APM (<i>FHL2</i> + <i>KLF14</i>)		31	0.777	0.604	0.576	9.30	0.000002	7.18

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

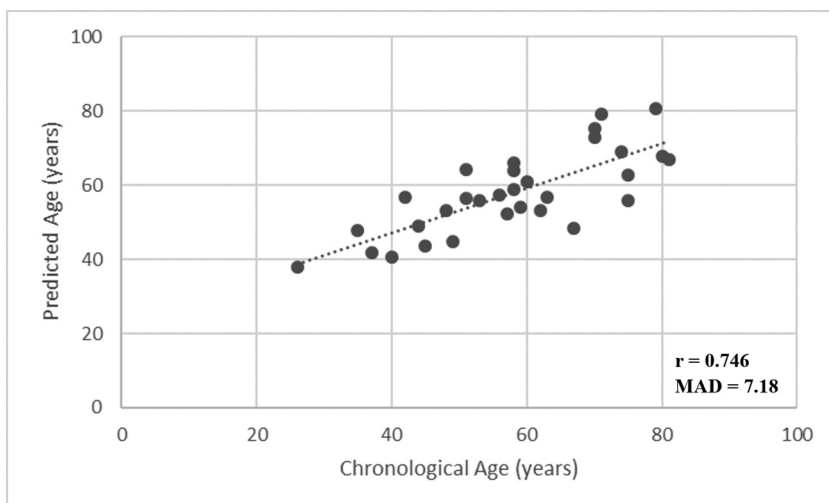


Figure 3. Predicted vs. chronological ages using the APM based on CpG sites at *FHL2* and *KLF14* genes built with methylation information from 31 bone samples obtained with the multiplex SNaPshot assay. Spearman correlation coefficient and MAD value from chronological age were plotted in the graph.

Table 4. Linear regression statistics at the 5 CpGs of the *ELOVL2*, *FHL2*, *MIR29B2C*, *KLF14* and *TRIM59* loci using methylation information from tooth samples obtained from a multiplex methylation SNaPshot analysis.

Locus	CpG Location	N	R	R ²	Corrected R ²	SE	P-value	MAD
Simple linear regression								
<i>ELOVL2</i>	Chr6:11044628	23	0.685	0.469	0.444	13.33	0.000311	11.27
<i>FHL2</i>	Chr2:105399282	23	0.331	0.110	0.067	17.43	0.122700	-
<i>KLF14</i>	Chr7:130734355	23	0.728	0.529	0.507	12.56	0.000084	9.68
<i>MIR29B2C</i>	Chr1:207823681	24	-0.080	0.006	-0.039	19.07	0.709684	-
<i>TRIM59</i>	Chr3:160450189	24	0.665	0.443	0.417	14.28	0.000389	11.51
Multiple linear regression								
APM (<i>ELOVL2</i> + <i>KLF14</i>)		23	0.886	0.785	0.764	12.56	2.09×10^{-7}	7.07

Abbreviations: N, number of samples; R, correlation coefficient; SE, standard error; MAD, mean absolute deviation (years) between chronological and predicted ages. Genomic positions were based on the GRCh38/hg38 assembly.

dual-locus model with the CpG sites at *ELOVL2* and *KLF14* in 23 tooth samples with positive PCR amplification for both genes. The multiple linear regression allowed us to obtain a strong age correlation value ($R = 0.886$), highly significant ($P = 2.09 \times 10^{-7}$), explaining 76.4% of age variance (Table 4). Predicting age of each individual through the formula $11.519 + 106.261 \times DNAm\ levels\ ELOVL2 + 291.877 \times DNAm\ levels\ KLF14$ (Supplemental Table S6), we estimated age with a correlation between predicted and chronological ages of 0.883 (Figure 4), and a MAD from chronological age of 7.07 years (Table 4). The threefold cross-validation made to test the accuracy of the model, allowed to estimate an averaged MAD from the chronological age for the three independent validation sets of 7.33 years, very close to the MAD of 7.07 from the total training data set. A validation by splitting the sample in two sets of 12 and 11 samples (training and validation sets) allowed to obtain an independent MAD value for the training set of 7.35 years. The model was applied to the validation set allowing to obtain a MAD of 6.34 years.



Figure 4. Predicted vs. chronological ages using the APM based on CpG sites at *ELOVL2* and *KLF14* genes built with methylation information from 23 tooth samples obtained with the multiplex SNaPshot assay. Spearman correlation coefficient and MAD value from chronological age were plotted in the graph.

Discussion

In the present study, we analysed bones and teeth from individuals of Portuguese ancestry through the bisulphite Sanger sequencing method to obtain methylation information of several CpGs located at *ELOVL2* (9 CpGs), *FHL2* (12 CpGs), *EDARADD* (4 CpGs), *PDE4C* (12 CpGs) and *MIR29B2C* (6 CpGs) genes. Sanger sequencing methodology, capturing several CpGs sites from the same locus, can allow the evaluation of DNA methylation levels from several positions. This can enable the finding of different age-correlated CpG sites and the development of APMs with more predictors and consequently with a more accurate age prediction. Moreover, we used the multiplex methylation SNaPshot assay of Jung et al.¹⁸ in the same tissues to analyse five CpG sites located at *ELOVL2*, *FHL2*, *KLF14*, *MIR29B2C* and *TRIM59*, repeatedly reported as age-associated genes. This methodology seems promising in forensic fields because of its capacity for multiplexing analysis, investigating simultaneously the DNA methylation level across several specific CpGs.

DNAm levels obtained from bones using bisulphite Sanger sequencing allowed the development of a final APM with six markers (*ELOVL2* CpG5, *ELOVL2* CpG6, *ELOVL2* CpG7, *MIR29B2C* CpG1, *EDARADD* CpG3 and *EDARADD* CpG4) revealing a high age correlation value, $R = 0.970$. The model showed high accuracy allowing to obtain a MAD from chronological age of 2.56 years, which suggest that this methodology could be suited for forensic purposes using bone-derived DNAm. The SNaPshot method allowed to develop from bones a final dual-locus APM with *FHL2* and *KLF14* genes showing a MAD from chronological age of 7.18 years, and a moderate correlation of 0.746 between predicted and chronological ages.

Bone samples were already used in studies by Gopalan et al.²² and Naue et al.¹³ for age-association methylation purposes. Naue et al.¹³ investigated through massive parallel sequencing whether 13 previously selected age-dependent loci have predictive value in

several forensically relevant tissues including bones. Using 29 deceased individuals (age range 0–87 years) they found no age-dependency for several genes, but a statistically significant age correlation (≥ 0.6) was observed for at least one CpG site within the amplified region of genes *ELOVL2*, *KLF14* and *TRIM59*. Gopalan et al.²² generated genome-wide DNA methylation data from 32 individual bone samples alongside with published data from 133 additional bone donors, both living and deceased spanning an age range of 49–112 years. The epigenome-wide association study on this combined dataset identify 108 sites of DNA methylation that showed significant relationship with age. The authors have developed an APM using lasso regression that produces highly accurate estimates of age. From all CpG sites that comprised the best model (37 bone clock CpG regions), two *ELOVL2* positions (cg16867657 and cg24724428) and one *KLF14* position (cg07955995) were included. In particular, the *ELOVL2* CpG6 (Chr6:11044644; cg16867657) from our study has already been included in Gopalan et al.²² and Naue et al.¹³ studies, revealing to be a promising marker for bone samples. This supports the idea that gene *ELOVL2* is a uniquely useful predictor of chronological age, as it is significantly associated with age across multiple tissue types.

In tooth samples, a moderate accurate model combining CpGs at *ELOVL2* and *KLF14* was obtained in our study with the SNaPshot method (MAD of 7.07 years). A similar value was obtained for dentin samples by Giuliani et al.²³ (MAD = 7.07 years) using 5 CpGs located at *ELOVL2*, *FHL2* and *PENK* genes addressed by Maldi-ToF mass spectrometry in 21 teeth extracted from living individuals with age ranging from 17 to 77 years. Bekaert et al.⁵ evaluating DNAm levels by pyrosequencing in 29 dentin samples of living individuals reported a multiple quadratic regression model developed with seven CpGs located at *PDE4C*, *ELOVL2* and *EDARADD* genes explaining 74% of age variance with a MAD of 4.86 years. The model included the *ELOVL2* CpG site (Chr6:11044628) selected in our study through the SNaPshot method. A recent study by Márquez-Ruiz et al.²⁴ testing methylation levels of specific CpG sites located in the *ELOVL2* and *PDE4C* genes by bisulphite pyrosequencing in 65 tooth samples from individuals aged 15–85 years old, developed an APM with nine CpG sites showing a mean absolute error (MAE) of 5.08 years. All these studies showed the usefulness of *ELOVL2* in development of epigenetic clocks using tooth samples. In contrast, the Sanger sequencing methodology in our study did not allow to develop a multi-locus APM in teeth, since only lower and moderate correlation values between age and DNAm were obtained. This suggests that this method could have limited usefulness for forensic age estimation using tooth samples. Meanwhile, a simple linear regression model using the high age-correlated *FHL2* CpG4 revealed a model accuracy of 11.35 years, which is a higher MAD value.

The present study has some limitations, being the sample size the major drawback. Moreover, despite we used 31 samples in both sets of bones and teeth not all of them were successfully amplified for all genes; only 29 fresh bone samples were addressed using the bisulphite Sanger sequencing methodology and 24 tooth samples using the multiplex SNaPshot assay. Larger sample sets and different loci could have higher statistical power and could be more representative of DNAm changes according to different age ranges and in different tissue types. We should take into account that also different CpGs and/or genes were addressed in both methodologies, consequently this could influence the accuracy of the developed models in each tissue type.

Conclusion

In conclusion, considering that to date only few reports used bone and tooth samples in development of models for forensic age estimation, we evaluated DNAm levels from bones and teeth using the bisulphite Sanger sequencing and SNaPshot methodologies. Our study allowed to develop a highly accurate APM in bone samples with six CpGs located at genes *ELOVL2*, *EDARADD* and *MIR29B2C* through bisulphite Sanger sequencing. The SNaPshot method allowed to construct two final dual-locus models, with *FHL2* and *KLF14* genes for bones and with *ELOVL2* and *KLF14* for teeth, exhibiting moderate age prediction accuracy. Our study suggests that skeletal and dental human remains, with high resistance to harsh conditions and often recoverable for long *post-mortem* intervals, can be prime targets for DNAm analyses in forensic contexts.

Acknowledgments

The authors would like to thank: Dr Cristina Cordeiro and Dr Carlos dos Santos that collected bone samples from deceased individuals in INMLCF; Dr Pedro Graça Pereira and Dr Dulce Madeira that collected bone and tooth samples from Bodies Donated to Science in *Departamento de Biomedicina, Unidade de Anatomia da Faculdade de Medicina da Universidade do Porto*; and Dr Armando Serra and Dr Lisa Sampaio. from *Serviço de Genética e Biologia Forenses* for technical support.

Disclosure statement

The authors declare that they have no conflict of interest.

Funding

This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) [FCT-PEst-OE/SADG/UI0283/2019]. H.D. has a PhD grant from FCT [SFRH/BD/117022/2016].

ORCID

Licínio Manco  <http://orcid.org/0000-0002-2636-0288>

References

1. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, Lynnerup N, Cattaneo C. The problem of aging human remains and living individuals: a review. *Forensic Sci Int.* 2009;193(1–3):1–13. doi:10.1016/j.forsciint.2009.09.008.
2. Freire-Aradas A, Phillips C, Lareu MV. Forensic individual age estimation with DNA: from initial approaches to methylation tests. *Forensic Sci Rev.* 2017;29(2):121–144.
3. Goel N, Karira P, Garg VK. Role of DNA methylation in human age prediction. *Mech Ageing Dev.* 2017;166:33–41. doi:10.1016/j.mad.2017.08.012.
4. Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples. *Forensic Sci Int Genet Supplement Series.* 2015;5:e144–e145. doi:10.1016/j.fsigss.2015.09.058.
5. Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics.* 2015;10(10):922–930. doi:10.1080/15592294.2015.1080413.

6. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Z, Pałeczka A, Kucharczyk K, Płoski R, Branicki W. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci Int Genet.* 2015;17:173–179. doi:10.1016/j.fsigen.2015.05.001.
7. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Z, Spas A, Parys-Proszek A, Kucharczyk K, Płoski R, Branicki W. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet.* 2015;14:161–167. doi:10.1016/j.fsigen.2014.10.002.
8. Eipel M, Mayer F, Arent T, Ferreira. MRP, Birkhofer C, Costa IG, Ritz-Timme S, Wagner W. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type - specific DNA methylation signatures. *Aging.* 2016;8:1034–1048. doi:10.18632/aging.100972.
9. Cho S, Jung S-E, Hong SR, Lee EH, Lee JH, Lee SD, Lee HY. Independent validation of DNA-based approaches for age prediction in blood. *Forensic Sci Int Genet.* 2017;29:250–256. doi:10.1016/j.fsigen.2017.04.020.
10. Correia Dias H, Cordeiro C, Corte Real F, Cunha E, Manco L. Age estimation based on DNA methylation using blood samples from deceased individuals. *J Forensic Sci.* 2020;65(2):465–470. doi:10.1111/1556-4029.14185.
11. Freire-Aradas A, Phillips C, Mosquera-Miguel A, Girón-Santamaría L, Gómez-Tato A, Casares de Cal M, Álvarez-Dios J, Ansedo-Bermejo J, Torres-Español M, Schneider PM, et al. Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci Int Genet.* 2016;24:65–74. doi:10.1016/j.fsigen.2016.06.005.
12. Naue J, Hoefsloot HCJ, Mook ORF, Rijlaarsdam-Hoekstra L, van der Zwalm MCH, Henneman P, Kloosterman AD, Verschure PJ. Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. *Forensic Sci Int Genet.* 2017;31:19–28. doi:10.1016/j.fsigen.2017.07.015.
13. Naue J, Sängner T, Hoefsloot HCJ, Lutz-Bonengel S, Kloosterman AD, Verschure PJ. Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. *Forensic Sci Int Genet.* 2018;36:152–159. doi:10.1016/j.fsigen.2018.07.007.
14. Aliferi A, Ballard D, Gallidabino MD, Thurtle H, Barron L, Syndercombe Court D. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Sci Int Genet.* 2018;37:215–226. doi:10.1016/j.fsigen.2018.09.003.
15. Lee HY, Jung S-E, Oh YN, Choi A, Yang WI, Shin K-J. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. *Forensic Sci Int Genet.* 2015;19:28–34. doi:10.1016/j.fsigen.2015.05.014.
16. Hong SR, Jung SE, Lee EH, Shin KJ, Yang WI, Lee HY. DNA methylation based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Forensic Sci Int Genet.* 2017;29:118–125. doi:10.1016/j.fsigen.2017.04.006.
17. Hong SR, Shin K-J, Jung S-E, Lee EH, Lee HY. Platform-independent models for age prediction using DNA methylation data. *Forensic Sci Int Genet.* 2019;38:39–47. doi:10.1016/j.fsigen.2018.10.005.
18. Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci Int Genet.* 2019;38:1–8. doi:10.1016/j.fsigen.2018.09.010.
19. Correia Dias H, Cordeiro C, Pereira J, Pinto C, Corte Real F, Cunha E, Manco L. DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay. *Forensic Sci Int.* 2020;311(110267):1–7.
20. Dmitrijeva M, Ossowski S, Serrano L, Schaefer MH. Tissue-specific DNA methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic Acids Res.* 2018;46(14):7022–7039. doi:10.1093/nar/gky498.

21. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin*. 2018;11(25):1–11. doi:[10.1186/s13072-018-0191-3](https://doi.org/10.1186/s13072-018-0191-3).
22. Gopalan S, Gaige J, Henn BM. DNA methylation-based forensic age estimation in human bone. *BioRxiv*. 2019. doi:[10.1101/801647](https://doi.org/10.1101/801647).
23. Giuliani C, Cilli E, Bacalini MG, Pirazzini C, Sazzini M, Gruppioni G, Franceschi C, Garagnani P, Luiselli D. Inferring chronological age from DNA methylation patterns of human teeth. *Am J Phys Anthropol*. 2016;159(4):585–595. doi:[10.1002/ajpa.22921](https://doi.org/10.1002/ajpa.22921).
24. Márquez-Ruiz AB, González-Herrera L, Luna JD, Valenzuela A. DNA methylation levels and telomere length in human teeth: usefulness for age estimation. *Int J Legal Med*. 2020;134(2):451–459. doi:[10.1007/s00414-019-02242-7](https://doi.org/10.1007/s00414-019-02242-7).



Age prediction in living: Forensic epigenetic age estimation based on blood samples

Helena Correia Dias^{a,b,c}, Eugénia Cunha^{b,c}, Francisco Corte Real^{c,d}, Licínio Manco^{a,*}

^a Research Centre for Anthropology and Health (CIAS), Department of Life Sciences, University of Coimbra, Portugal

^b Centre for Functional Ecology (CEF), Laboratory of Forensic Anthropology, Department of Life Sciences, University of Coimbra, Portugal

^c National Institute of Legal Medicine and Forensic Sciences, Portugal

^d Faculty of Medicine, University of Coimbra, Portugal

ARTICLE INFO

Keywords:

Forensic sciences
Forensic epigenetics
Age the living
DNA methylation age
CpGs

ABSTRACT

DNA methylation analysis in a variety of genes has brought promising results in age estimation. The main aim of this study was to evaluate DNA methylation levels from four age-correlated genes, *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, in blood samples of healthy Portuguese individuals. Fifty-three samples were analyzed through the bisulfite polymerase chain reaction (PCR) sequencing method for CpG dinucleotide methylation status. Linear regression models were used to analyze relationships between methylation levels and chronological age. The highest age-associated CpG in each locus was chosen to build a multi-locus age prediction model (APM), allowing to obtain a Mean Absolute Deviation (MAD) between chronological and predicted ages of 5.35 years, explaining 94.1% of age variation. Validation approaches demonstrated the accuracy and reproducibility of the proposed multi-locus APM. Testing the APM in 51 blood samples from deceased individuals a MAD of 9.72 years was obtained. Potential differences in methylation status between samples from living and deceased individuals could exist since the highest age-correlated CpGs were different in some genes between both groups. In conclusion, our study using the bisulfite PCR sequencing method is in accordance with the high age prediction accuracy of DNA methylation levels in four previously reported age-associated genes. DNA methylation pattern differences between blood samples from living and deceased individuals should be taken into account in forensic contexts.

1. Introduction

Forensic age estimation is an important clue for identification purposes. In cases of living individuals, forensic age estimation could be essential in judicial, criminal or civil situations, including cases of immigration or refugees (where the identity and age of individuals are unclear), cases of minors (in questions related to imputability), for determination of criminal responsibility, or even in civil cases of pensionable age (for old adults lacking documents) [1–3].

Biochemical and genetic age predictive biomarkers used until now, such as accumulation of D-aspartic acid in proteins [4], shortening of telomeres, or deletion of mitochondrial DNA [5], have shown low accuracy and other inconsistencies, being considered inappropriate for forensic and identity science casework [6,7]. Consequently, there is still no standard method for forensic age estimation. In this scenario, the evaluation of DNA methylation has been gaining relevance in age estimation investigations because it has brought the most promising

results [8,9].

Although influenced by genetic, environmental, disease and stochastic factors [10–13], there is growing evidence that DNA methylation patterns of specific CpG sites in genes such as *ELOVL2*, *FHL2*, *EDARADD*, *ASPA*, *PDE4C*, *PENK*, *C1orf132*, *TRIM59* and *KLF14* are associated with chronological age in several tissues such as blood, buccal swabs and teeth, for developing several forensic highly accurate age-prediction models (APMs) [13–28]. Most studies used bisulfite modification and pyrosequencing to assess the methylation patterns of CpG sites in blood samples for implementation of APMs [15,16,18,19,23,27,28]. Meanwhile, the bisulfite polymerase chain reaction (PCR) sequencing methodology was shown to be an efficient and economical alternative tool for rapid quantification of DNA methylation, with similar linearity and accuracy than pyrosequencing analysis [29,30].

In a previous study, our group have used the bisulfite-PCR sequencing methodology for evaluation the correlation between

* Corresponding author at: Research Centre for Anthropology and Health (CIAS), Department of Life Sciences, University of Coimbra, Rua do Arco da Traição, Coimbra 3000-456 Portugal.

E-mail address: lmanco@antrop.uc.pt (L. Manco).

<https://doi.org/10.1016/j.legalmed.2020.101763>

Received 4 May 2020; Received in revised form 26 June 2020; Accepted 16 July 2020

Available online 21 July 2020

1344-6223/ © 2020 Elsevier B.V. All rights reserved.

chronological age and DNA methylation levels located at *ELOVL2*, *FHL2*, *EDARADD*, *PDE4C* and *C1orf132* genes in 51 blood samples from deceased individuals [24]. The combination of the five strongest age-correlation markers from each gene in a final APM seems to be informative and could have potential application in forensic analysis [24]. In the present study, using blood samples of living individuals, we applied the same bisulfite-PCR sequencing method to evaluate CpG methylation levels in the four age-correlated genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*. The main purposes of the current study were: (i) to develop a specific APM for blood samples from living individuals of Portuguese ancestry, and (ii) to test the accuracy of this APM in blood samples from deceased individuals.

2. Material and methods

2.1. Study population

Peripheral blood samples of 53 healthy individuals of Portuguese ancestry (35 females, 18 males; aged 1–95 years), were collected in EDTA-tubes from users of *Biobanco - Hospital Pediátrico de Coimbra* and other hospitals for developing our DNA methylation model. An independent set of blood samples of 18 healthy individuals (10 females and 8 males; aged 1–93 years old) was used for validation the developed APM. Additionally, 51 blood samples from deceased individuals (7 females and 44 males; aged 24–86 years old), previously reported in a study by our group [24], were used for testing the reproducibility of the developed APM.

The study protocol was approved by the ethical Committee of *Faculdade de Medicina da Universidade de Coimbra* (n° 038-CE-2017). Written informed consent was previously obtained from adult participants and from children's parents under the age of 18 years.

2.2. DNA methylation analysis by bisulfite conversion and PCR-sequencing

Genomic DNA was extracted with the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and quantified in a Nanodrop spectrophotometer (Thermo Fisher Scientific). Genomic DNA was subjected to bisulfite conversion using the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, USA) according to the instructions of manufacturer, as previously described [24]. After conversion, modified DNA samples were submitted to PCR amplification for selected regions of genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*, followed by Sanger sequencing, as previously described [24]. To assess the reproducibility of the method, two separate PCR amplifications and Sanger sequencing analyses were performed in about 10% of DNA samples for all genes.

2.3. Methylation quantification of DNA sequencing data

The methylation status of cytosine (C) in each CpG dinucleotides was estimated according to Jiang *et al.* [29] and Parrish *et al.* [30], by measuring the ratio between peak height values of C and thymine (T) through the formula $[C/C + T]$ in the sequencing chromatogram extracted from Chromas (Version 2.32, Technelysium). In each CpG, a single C reveal complete methylation (100%), a single T complete unmethylation (0%) and overlapping C and T partial methylation (0–100%). As the reverse primer was used in the sequencing reaction, the reverse-complement strand of the sequencing chromatogram was used to estimate the ratio between peak heights of C and T. About 50% of the samples included in this study was evaluated by two independent investigators to indemnify the inter-observatory error.

2.4. DNA methylation standards

Each primer set used for bisulfite sequencing was independently verified, to confirm the accuracy of sequencing data using the DNA methylation commercial standards EpiTect Control DNA, methylated

and unmethylated (Qiagen, Hilden, Germany). Standard DNA samples premixed at methylation levels of 0%, 50%, and 100% were used for analysis (Supplementary Fig. S1).

2.5. Statistical analyses

Statistical analyses were performed using IBM SPSS statistics software for Windows, version 24.0 (IBM Corporation, Armonk, NY, USA). Simple linear regressions were used to analyze relationships between methylation levels and chronological age at each single CpG site. The CpG site showing the highest age-correlation from each gene was selected for combined analysis using multiple linear regression. Using the regression coefficients, we predicted age of individuals applying the multiple linear regression formula: $Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Nx_N$, where Y is the age of the individual, b_0 is the intercept, b_1 , b_2 , b_N the slope of the selected CpGs and x_1 , x_2 , x_N the methylation values of the selected CpGs.

The Mean Absolute Deviation (MAD) between predicted and chronological ages was calculated for the training sample (53 individuals) and for subsets of four distinct age categories: < 18 years, 19–39 years, 40–60 years and > 61 years. Each obtained result was interpreted as either correct or incorrect if the predicted age was concordant with the chronological age using a cutoff value according to the standard error (SE) of estimate obtained in the developed APM.

An independent set of 18 blood samples from living individuals was used for validation of the developed APM. In addition, a 4-fold cross validation was performed in the training set of 53 living individuals, which consisted in removing randomly a set of samples from the training set and to develop four independent multiple linear regressions on the remaining samples. The removed samples were assigned as validation sets to calculate the cross validation mean MAD values. An additional validation was performed by splitting the complete training set of 53 living individuals into 2 subsets of 27 and 26 blood samples (training and validation sets). An independent regression was developed for the training set and applied to the validation set.

The evaluation of differences between gender was made through comparison of two regression lines relating chronological age and DNA methylation levels of each gene at two levels (males/females) of the categorical factor, using the software STATGRAPHICS Centurion XV, version 15.2.05 (StatPoint Technologies, Inc., VA).

3. Results

In the present study, we evaluated DNA methylation levels of several CpG sites located at *ELOVL2* (9 CpGs), *FHL2* (12 CpGs), *EDARADD* (4 CpGs) and *PDE4C* (12 CpGs) genes through bisulfite conversion followed by PCR and direct Sanger sequencing.

The reproducibility of the direct bisulfite Sanger sequencing was made in about 10% of the samples by two separate PCR amplifications and sequencing analyses. The mean percentage difference in DNA methylation levels for all CpGs in each gene was 6.5% for *PDE4C* (5 samples), 4.3% for *EDARADD* (7 samples), 2.9% for *ELOVL2* (5 samples), and 3.9% for *FHL2* (7 samples).

The accuracy of methylation levels obtained by bisulfite sequencing, was evaluated by analyzing the PCR mixture amplification for each locus using three different methylation rates of 0%, 50%, and 100% (Supplementary Fig. S1). Bisulfite sequencing resulted in DNA methylation levels that bore a significant linear relationship to expected methylation levels.

No statistical significant differences were observed between males and females in the overall training sample (53 individuals) comparing slopes and intercepts of two simple linear regression lines of methylation status and age allowing to ignore gender differences in subsequent analyses (data not shown).

Table 1

Linear regression statistics of the best age predictors in *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C* genes testing for the association between CpG sites and chronological age in 53 blood samples of living individuals.

Locus	CpG site	Location	R	Corrected R ²	SE	P-value	MAD
Simple linear regression							
<i>ELOVL2</i>	CpG6	Chr6:11,044,644	0.936	0.874	10.41	7.97×10^{-25}	8.01
<i>FHL2</i>	CpG3	Chr2:105,399,291	0.940	0.881	10.11	1.78×10^{-25}	7.81
<i>EDARADD</i>	CpG3	Chr1: 236,394,382	-0.888	0.784	13.63	7.86×10^{-19}	10.57
<i>PDE4C</i>	CpG2	Chr19: 18,233,133	0.852	0.720	15.53	6.32×10^{-16}	11.87
Multiple linear regression							
APM			0.972	0.941	7.13	1.11×10^{-29}	5.35

Abbreviations: R, Regression coefficient; SE, standard error; MAD, Mean Absolute Deviation from the chronological age; APM, Age Prediction Model. Location of CpGs is according to the human GRCh38/hg38 assembly.

3.1. Development of an age prediction model in a training set of 53 living Portuguese individuals

Simple linear regressions testing the correlation between methylation levels and chronological age revealed significant associations ($P < 0.05$) for all CpGs from the genes *ELOVL2* and *FHL2* (except CpG7, CpG9 and CpG10) (Supplementary Table S1). Significant associations with methylation levels were also observed for *EDARADD* CpG2, CpG3 and CpG4, as well as for *PDE4C* CpG1 to CpG6 and CpG9 (Supplementary Table S1). The strongest age-correlations within each gene were observed for: *ELOVL2* CpG6 ($R = 0.936$; $P = 7.97 \times 10^{-25}$), explaining 87.4% of variation in age; *FHL2* CpG3 ($R = 0.940$; $P = 1.78 \times 10^{-25}$), explaining 88.1% of variation in age; *EDARADD* CpG3 ($R = -0.888$; $P = 7.86 \times 10^{-19}$), explaining 78.4% of variation in age; and *PDE4C* CpG2 ($R = 0.852$; $P = 6.32 \times 10^{-16}$), explaining 72% of variation in age (Table 1; Fig. 1). The predicted age of individuals was calculated through the simple linear regression coefficients for the individual strongest age-associated markers and the obtained MAD values were as follows: 8.01 years for *ELOVL2* CpG6; 7.81 years for *FHL2* CpG3; 10.57 years for *EDARADD* CpG3; and 11.87 for *PDE4C* CpG2 (Table 1).

Combining the methylation information of CpGs most highly associated with age per locus, *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2, the final multi-locus APM reveals a higher age correlation value ($R = 0.972$), highly significant ($P = 1.11 \times 10^{-29}$), explaining 94.1% of the variation in age (adjusted $R^2 = 0.941$) (Table 1). Predicted age of individuals was calculated through the multiple linear regression coefficients using the formula $(-81.879) + 103.031 \times \text{DNA methylation levels } ELOVL2 \text{ CpG6} + 99.331 \times \text{DNA methylation levels } FHL2 \text{ CpG3} - 58.97 \times \text{DNA methylation levels } EDARADD \text{ CpG3} + 35.843 \times \text{DNA methylation levels } PDE4C \text{ CpG2}$. A strong correlation between predicted and chronological ages was obtained (Spearman correlation coefficient, $r = 0.972$) with a MAD from chronological age of 5.35 years (Fig. 2). Correct predictions were 75.5% assuming that chronological and predicted ages match ± 7 years (Table 2).

3.2. Differences between predicted and chronological ages with aging

Larger differences between predicted and chronological ages were observed with increasing of age (Fig. 3). To investigate these age-related differences, we divided our training set of 53 living individuals in four age groups (< 18 years old; 18–39 years old; 40–60 years old; > 61 years old) to estimate MAD and percentage of correct predictions in each age range group. MAD value is higher in the two older age categories (3 and 4 age range group): 40–60 years old (MAD = 6.49 years) and > 61 years old (MAD = 6.27 years). In concordance, the lower percentage of correct predictions was observed in the older age groups 3 and 4 (58.3% and 76.5%, respectively). For younger individuals < 18 years old (MAD = 3.26 years) and age range group 19–39 years old (MAD = 4.99 years), the smaller MAD values were obtained and

higher values of correct predictions were observed (83.3% in both age groups) (Table 2).

3.3. Testing the accuracy of the developed APM

3.3.1. Living individuals

For evaluation of the accuracy of our multi-locus APM with *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 markers we made a 4-fold cross validation using our data set of 53 blood samples. The mean MAD value obtained amongst the four validation sets was 6.20 years, similar to the obtained in the overall population (MAD = 5.35 years). The validation approach through splitting the overall sample set of 53 living in two sets of 27 and 26 samples (training and validation sets) allowed to obtain a MAD value of 6.08 years in the new training set of 27 samples. Applying this multiple model on the validation set of 26 samples, a MAD of 5.81 years was obtained. Both independent MAD values were very close to the MAD of 5.35 years for the overall sample. Additionally, using an independent sample set of 18 blood samples from healthy Portuguese individuals, the performance of our multi-locus APM developed in the overall sample of 53 living individuals was evaluated. Based on the multiple linear regression model a strong correlation between predicted and chronological ages was observed (Spearman correlation coefficient, $r = 0.977$), with a MAD from the chronological age of 4.98 years (Supplementary Fig. S2).

3.3.2. Deceased individuals

We tested the reproducibility of the developed multi-locus APM in a different set of 51 blood samples from deceased individuals for which methylation data was previously addressed [24]. The obtained correlation value between predicted and chronological ages was 0.792, with a MAD value from chronological age of 9.72 years (Supplementary Fig. S3).

Some differences were observed in DNA methylation levels of blood samples between living and deceased individuals. The correlation values obtained in living individuals for the selected sites chosen for development of the final APM (*ELOVL2* CpG6, *EDARADD* CpG3, *FHL2* CpG3 and *PDE4C* CpG2) showed strong or very strong age correlation values ($0.852 < R < 0.940$), whereas in deceased individuals we observed for the same sites lower age correlation values ($0.459 < R < 0.764$). In concordance we observed a decrease of model accuracy in blood samples from deceased individuals when applying the APM previous developed in living individuals (MAD = 5.35 vs. 9.72). In blood samples from deceased individuals [24], the highest age-correlation site selected from two genes (*ELOVL2* CpG4 and *FHL2* CpG2) is different from that selected in this study using blood samples from living individuals. Moreover, the degradation of methylation at *postmortem* is observed among the general CpG sites from all genes: $R < 0.936$ in living vs. $R < 0.785$ in deceased for *ELOVL2*; $R < 0.940$ in living vs. $R < 0.465$ in deceased for *FHL2*; $R < 0.888$ in living vs. $R < 0.621$ in deceased for *EDARADD*; and $R < 0.852$ in living vs. $R < 0.592$ in deceased for *PDE4C*.

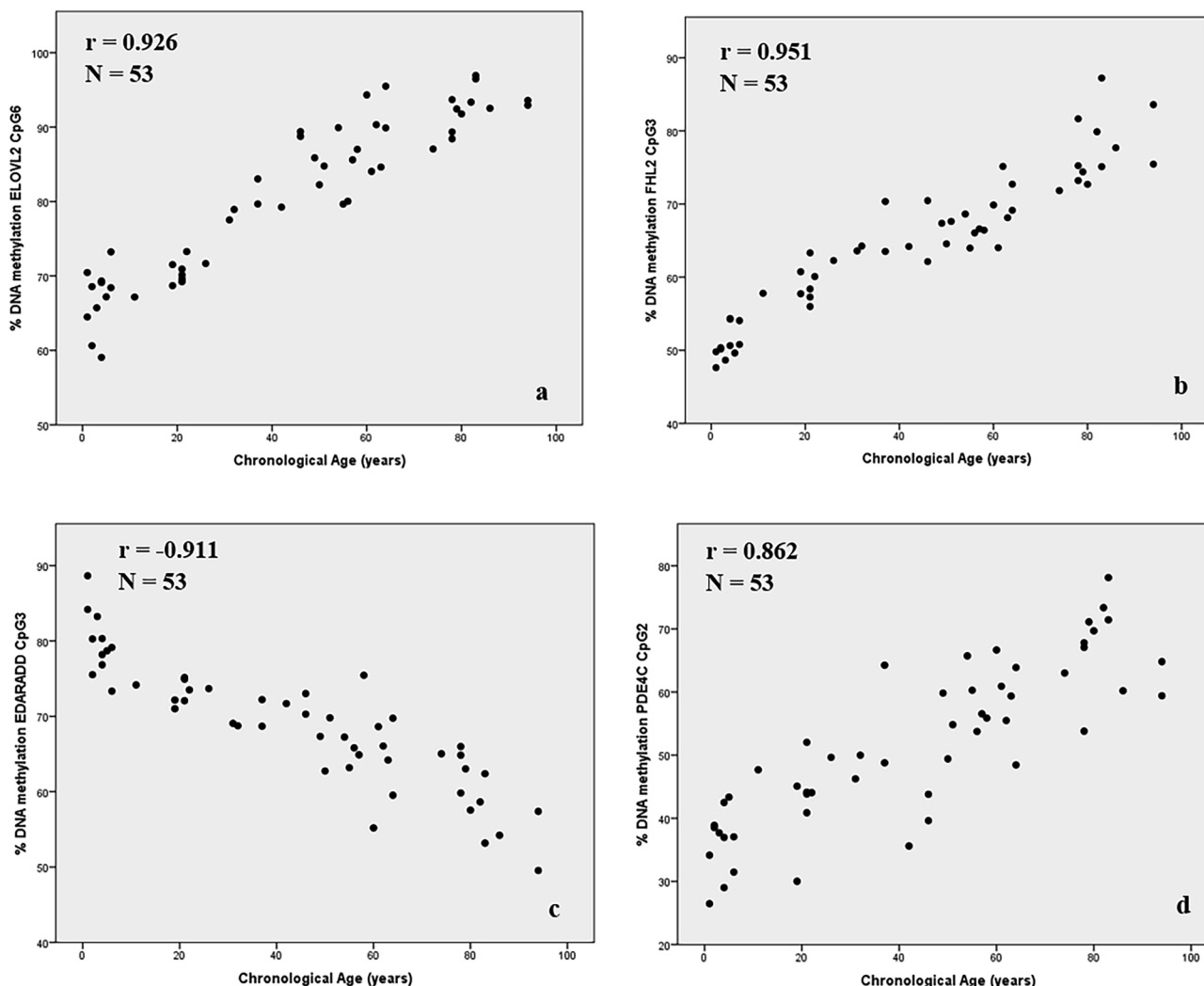


Fig. 1. Correlation between methylation levels and chronological age for CpG6 of *ELOVL2* (a), CpG3 of *FHL2* (b), CpG3 of *EDARADD* (c) and CpG2 of *PDE4C* (d) in 53 healthy individuals. The corresponding Spearman correlation coefficients (r) and sample sizes (N) are depicted inside each plot.

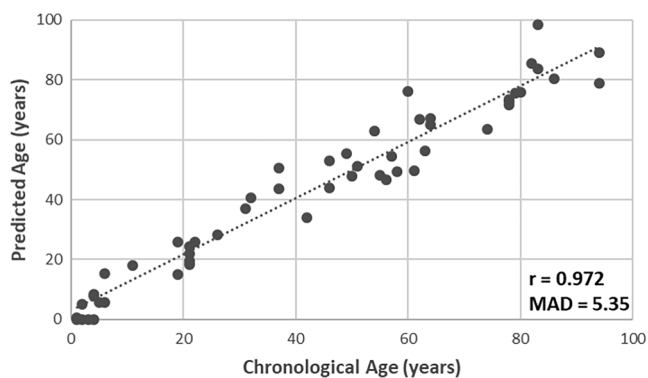


Fig. 2. Plot with predicted age (years) versus chronological age (years) of the 53 individuals using the developed APM based on *ELOVL2* CpG6, *FHL2* CpG3, *EDARADD* CpG3 and *PDE4C* CpG2 markers. Spearman correlation coefficient was 0.972.

4. Discussion

Age estimation plays a relevant role in forensic science since it can be very useful in criminal, legal and civil investigations, including for the establishment of criminal responsibility or in immigration cases [3]. In this work, we used the bisulfite PCR-sequencing method to analyze

Table 2

Mean absolute deviation (MAD) between predicted and chronological ages stratified according age groups in the training set.

Group	Age range (years)	N	MAD	Correct Predictions (%)
1	< 18	12	3.26	83.3
2	19–39	12	4.99	83.3
3	40–60	12	6.49	58.3
4	> 61	17	6.27	76.5
Total	1–95	53	5.35	75.5

Abbreviations: N, number of samples. Correct predictions were estimate according to the standard error ($SE = 7.13$) for the developed age prediction model.

the methylation patterns of CpG markers in four well known age-associated loci (*ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*) selected based on their powerful value in several previous studies [13–28].

The training set of 53 blood samples from healthy Portuguese individuals revealed the highest age-correlation value per locus for *ELOVL2* CpG6, *EDARADD* CpG3, *FHL2* CpG3 and *PDE4C* CpG2. Evaluating simultaneously the methylation information of these sites in a final four-locus APM, a good accuracy was revealed with a MAD from chronological age of 5.35 years. This MAD value is similar to the obtained in other APMs using the pyrosequencing methodology which has

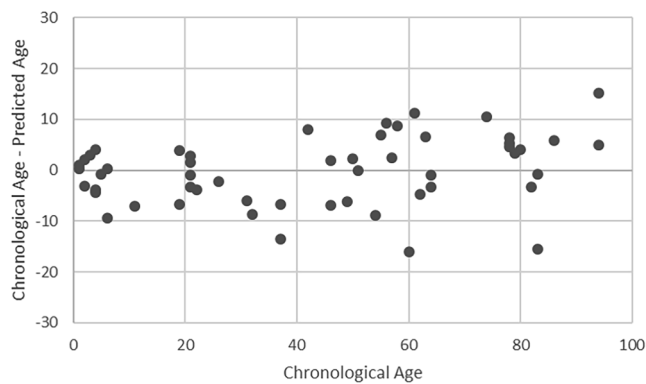


Fig. 3. Differences between chronological and predicted age plotted against chronological age.

been the preferred method by forensic epigeneticists to assess site specific DNA methylation levels in age predictions [15,16,18,19].

Meanwhile, the bisulfite Sanger sequencing methodology has shown promising results in blood samples from deceased individuals [24]. In this previous study, we have investigated the same four genes (*ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*) in addition to *C1orf132*, to develop an APM using the highest age-correlated CpG from each locus, revealing a correlation coefficient of 0.888, explaining 76.3% of age variation and showing a MAD of 6.08 years [24].

Comparing both studies, potential differences in methylation status between samples from living and deceased individuals could exist since the highest age-correlated CpGs were different in some genes between both groups. Moreover, the correlation between DNA methylation and age obtained in genes *ELOVL2*, *EDARADD*, *FHL2* and *PDE4C* is lower in blood samples of deceased individuals vs. living individuals, with *ELOVL2* revealing the higher and most similar DNAm levels in both groups. In concordance, when we applied the final APM built for living individuals to the independent set of 51 blood samples from deceased individuals [24], we obtained a higher value of MAD (9.72 years), which represents a decrease of the model accuracy in this set of samples. This data could be explained by *postmortem* changes that can alter the methylation status among specific loci, although this issue has not yet been clarified until now [31]. DNA methylation changes after death suggest that a specific APM developed for blood samples from living individuals cannot accurately be applied for blood samples from deceased individuals, in concordance with data obtained in our previous study with the SNaPshot methodology [25]. *ELOVL2* gene is an exception in concordance with general data recognizing this locus as a stable age-associated marker exhibiting consistent age-related changes across different tissues such as blood, teeth or saliva [15,21,32,33]. This explains the potential power of *ELOVL2* gene for age estimations in forensic contexts proposed by several studies [13,18,20,32,34].

The accuracy obtained in independent models for blood samples of living (this study) and deceased individuals [24] using the bisulfite PCR-sequencing methodology was similar (MAD values of 5.35 and 6.08 years, respectively), attesting the reproducibility of the bisulfite PCR-sequencing methodology. This approach is simple, does not include complex procedures, being less time-consuming and less expensive, showing to have near equivalent accuracy to pyrosequencing [29,30]. Moreover, bisulfite PCR-sequencing enable us to assess the percent methylation across multiple CpG sites allowing the possibility to choose several CpGs of interest.

In the present study, larger differences with the increase of age were observed in the training set of 53 living individuals, in concordance with previous studies [16,18,19,25,31]. Grouping the samples into four age categories, older individuals showed an increased MAD, with a lower percentage of correct predictions. The fact that DNA methylation patterns predict age with more accuracy in younger than in older

individuals suggests increased inter-individual variation within older people, possibly due to environmental, diseases and stochastic factors [10–13].

Another relevant issue is the possible influence of sex in DNA methylation levels of age correlated markers. To date, there is no consensus for a relationship between age-associated DNA methylation levels and sex [16,19,25,32,35,36]. Similar to other studies [16,25,35,36] our results showed no significant differences between males and females, suggesting that gender has no influence in DNA methylation levels of age-related CpG sites located at *ELOVL2*, *EDARADD*, *FHL2* and *PDE4C* genes.

The proposed APM developed using 53 blood samples from living individuals was tested in an independent set of 18 blood samples, also from living individuals, revealing reliability and reproducibility (MAD = 4.98 years). Moreover, two additional validation approaches have been made revealing high reproducibility. Through 4-fold cross validation, using the training set of 53 living individuals, a mean MAD value of 6.20 years was obtained for the validation sets; and through splitting the overall sample set of 53 living individuals in two sets with similar size (training and validation sets), a MAD value of 6.08 years was obtained for the training set and a MAD of 5.81 years for the validation set.

Moreover, we can hypothesized that our developed APM for blood samples from living individuals can be applied to fresh bloodstains in forensic contexts due to DNA methylation stability, as it was demonstrated for developed models in two previous studies [18,36]. Huang *et al.* [36] observed no statistically significant differences in age prediction between blood samples and bloodstains, as well as in the comparison of predicted age on the basis of different period bloodstains. These results were similar to the observed by Zbieć-Piekarska [18] in which the rate of corrected predictions in bloodstains seems to be not related with the time of storage. Based on this, it is also possible that our model developed for blood samples from living individuals could be applied to forensic casework in cases of bloodstains from unknown origin.

As conclusion, using a training set of 53 blood samples of living individuals, we developed a final APM with the highly age-associated CpGs in genes *ELOVL2*, *FHL2*, *EDARADD* and *PDE4C*. The model revealed an accurate age estimation with a MAD from chronological age of 5.35 years. Our results through the bisulfite PCR-sequencing methodology are in concordance with previous studies using the pyrosequencing assay, revealing being accurate for age estimations in forensic casework, at least in blood. Moreover, a similar accuracy was observed comparing with the APM developed for blood samples from deceased individuals, showing reproducibility and applicability of bisulfite PCR-sequencing for age estimation in forensic contexts using blood. Differences in DNA methylation levels between living and deceased individuals suggest that DNA methylation *postmortem* changes can occur. This hypothesis reveal the necessity and usefulness of development of APMs specific not only for each type of sample but also considering the safety of the donor.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) (FCT-PEst-OE/SADG/UI0283/2019). H.D. has a PhD grant from FCT (SFRH/BD/117022/2016). The authors thank all the supporting for the present studying included the people who collected blood samples from living individuals in *Centro Hospitalar e Universitário*

de Coimbra and deceased individuals in *INMLCF (Instituto Nacional de Medicina Legal e Ciências Forenses)*.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.legalmed.2020.101763>.

References

- [1] E. Cunha, E. Baccino, L. Martrille, F. Ramsthaler, J. Prieto, Y. Schuliar, N. Lynnerup, C. Cattaneo, The problem of aging human remains and living individuals, *Forensic Sci. Int.* 193 (2009) 1–13.
- [2] E. Baccino, E. Cunha, C. Cattaneo, Aging the dead and the living, *Elsevier*, 2013, pp. 42–48.
- [3] W. Parson, Age estimation with DNA: from forensic DNA Fingerprinting to Forensic (Epi) genomics: a mini-review, *Gerontology* (2018), <https://doi.org/10.1159/000486239>.
- [4] R.C. Dobberstein, J. Huppertz, N. von Wurmb-Schwark, S. Ritz-Timme, Degradation of biomolecules in artificially and naturally aged teeth: implications for age estimation based on aspartic acid racemization and DNA analysis, *Forensic Sci. Int.* 179 (2008) 181–191.
- [5] A. Tsuji, A. Ishiko, T. Takasaki, N. Ikeda, Estimating age of humans based on telomere shortening, *Forensic Sci. Int.* 126 (2002) 197–199.
- [6] C. Meissner, S. Ritz-Timme, Molecular pathology and age estimation, *Forensic Sci. Int.* 203 (2010) 34–43.
- [7] A.M. Albert, C.L. Wright, DNA prediction in Forensic Anthropology and the Identity Sciences, *G. J. Anthropol. Res.* 2 (2015) 1–6.
- [8] S.E. Jung, K.J. Shin, H.Y. Lee, DNA methylation-based age prediction from various tissues and body fluids, *BMB. Rep.* 50 (2017) 546–553.
- [9] N. Goel, P. Karira, V.K. Garg, Role of DNA methylation in human age prediction, *Mech. Ageing Dev.* 166 (2017) 33–41.
- [10] M.P. Boks, E.M. Derks, D.J. Weisenberger, E. Strengman, E. Janson, I.E. Sommer, R.S. Khan, R.A. Ophoff, The relationship of DNA methylation with age, gender and genotype in twins and healthy controls, *PLoS ONE* 4 (2009) e6767.
- [11] H. Heyn, N. Li, H.J. Ferreira, S. Moran, D.G. Pisano, A. Gomez, J. Diez, J.V. Sanchez-Mut, F. Setien, F.J. Carmona, A.A. Puca, S. Sayols, M.A. Pujana, J. Serra-Musach, I. Iglesias-Platas, F. Formiga, A.F. Fernandez, M.F. Fraga, S.C. Heath, A. Valencia, I.-G. Gut, J. Wang, M. Esteller, Distinct DNA methylomes of newborns and centenarians, *PNAS* 109 (2012) 10522–10527.
- [12] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, *Nat. Genet.* 33 (2003) 245–254.
- [13] M. Spólnicka, M. Pośpiech, E. Peplowska, B. Zbieć-Piekarska, R. Makowska, Z. Pięta, A. Karłowska-Pik, J. Ziemkiewicz, B. Weżyk, M. Gasperowicz, P. Bednarczuk, T. Barcikowska, M. Żekanowski, C. Płoski, R. Wojciech Branicki, DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups, *Int. J. Legal Med.* 132 (2018) 1–11.
- [14] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, A.M. Di Blasio, D. Gentilini, G. Vitale, S. Collino, S. Rezzi, G. Castellani, M. Capri, S. Salvioli, C. Franceschi, Methylation of ELOVL2 gene as a new epigenetic marker of age, *Aging Cell* 11 (2012) 1132–1134.
- [15] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van de Voorde, R. Decorte, A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples, *Forensic Sci. Int. Genet. Supplement Series* 5 (2015) e144–e145.
- [16] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van de Voorde, R. Decorte, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, *Epigenetics* 10 (2015) 922–930.
- [17] C. Giuliani, E. Cilli, M.G. Bacalini, C. Pirazzini, M. Sazzini, G. Gruppioni, C. Franceschi, P. Garagnani, D. Luiselli, Inferring Chronological Age from DNA Methylation Patterns of Human Teeth, *Am. J. Phys. Anthropol.* 159 (2016) 585–595.
- [18] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, Z. Makowska, A. Spas, A. Parys-Proszek, K. Kucharczyk, R. Płoski, W. Branicki, Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science, *Forensic Sci. Int. Genet.* 14 (2015) 161–167.
- [19] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Z. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki, Development of a forensically useful age prediction method based on DNA methylation analysis, *Forensic Sci. Int. Genet.* 17 (2015) 173–179.
- [20] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. Henneman, A.D. Kloosterman, P.J. Verschure, Chronological Age Prediction based on DNA Methylation: Massive Parallel Sequencing and Random Forest Regression, *Forensic Sci. Int. Genet.* 31 (2017) 19–28.
- [21] S.E. Jung, S.M. Lim, S.R. Hong, E.H. Lee, K.J. Shin, H.Y. Lee, DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples, *Forensic Sci. Int. Genet.* 38 (2019) 1–8.
- [22] C. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, D.O. Bauerschlag, K.H. Jöckel, R. Erbel, T.W. Mühleisen, M. Zenke, T.H. Brümmerdorf, W. Wagner, Aging of blood can be tracked by DNA methylation changes at just three CpG sites, *Genome Biol.* 15 (2014) R24.
- [23] J.L. Park, J.H. Kim, E. Seo, D.H. Bae, S.Y. Kim, H.C. Lee, K.M. Woo, Y.S. Kim, Identification and evaluation of age-correlated DNA methylation markers for forensic use, *Forensic Sci. Int. Genet.* 23 (2016) 64–70.
- [24] H. Correia Dias, C. Cordeiro, F. Corte Real, E. Cunha, L. Manco, Age estimation based on DNA methylation using blood samples from deceased individuals, *J. Forensic Sci.* (2019) 1–6.
- [25] H. Correia Dias, C. Cordeiro, J. Pereira, C. Pinto, F. Corte Real, E. Cunha, L. Manco, DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay, *Forensic Sci. Int.* 311 (2020) (110267) 1–7.
- [26] L. Kananen, S. Marttila, T. Nevalainen, J. Jylhävä, N. Mononen, M. Kähönen, O.T. Raitakari, T. Lehtimäki, M. Hurme, Aging-associated DNA methylation changes in middle-aged individuals: the Young Finns study, *BMC Genomics* 17 (2016) 103.
- [27] J. Fleckhaus, A. Freire-Aradas, M.A. Rothschild, P.M. Schneide, Impact of genetic ancestry on chronological age prediction using DNA methylation analysis, *Forensic Sci. Int. Genet. Supplement Series* 6 (2017) e399–e400.
- [28] Z. Thong, X.L.S. Chan, J.Y.Y. Tan, E.S. Loo, C.K.C. Syn, Evaluation of DNA methylation-based age prediction on blood, *Forensic Sci. Int. Genet. Supplement Series* 6 (2017) e249–e251.
- [29] M. Jiang, Y. Zhang, J. Fei, X. Chang, W. Fan, X. Qian, T. Zhang, D. Lu, Rapid quantification of DNA methylation by measuring relative peak heights in direct bisulfite-PCR sequencing traces, *Lab. Investig.* 90 (2010) 282–290.
- [30] R.R. Parrish, J.J. Day, F.D. Lubin, Direct bisulfite sequencing for examination of DNA methylation patterns with gene and nucleotide resolution from brain tissues, *Curr Protoc. Neurosci.* 7 (2012) Unit 7.24.
- [31] Y. Hamano, S. Manabe, C. Morimoto, S. Fujimoto, M. Ozeki, K. Tamaki, Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting, *Leg. Med.* 21 (2016) 5–10.
- [32] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, K. Zhang, Genome-wide methylation profiles reveal quantitative views of human aging rates, *Mol. Cell* 49 (2013) 359–367.
- [33] J. Naue, T. Sängler, H.C.J. Hoefsloot, S. Lutz-Bonengel, A.D. Kloosterman, P.J. Verschure, Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing, *Forensic Sci. Int. Genet.* 36 (2018) 152–159.
- [34] A. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, *PLoS ONE* 8 (2013) e67378.
- [35] A. Freire-Aradas, C. Phillips, L. Girón-Santamaría, A. Mosquera-Miguel, A. Gómez-Tato, M.Á.C. de Cal, J. Álvarez-Dios, M.V. Lareu, Tracking age correlated DNA methylation markers in the young, *Forensic Sci. Int. Genet.* 36 (2018) 50–59.
- [36] Y. Huang, J. Yan, J. Hou, X. Fu, L. Li, Y. Hou, Developing a DNA methylation assay for human age prediction in blood and bloodstain, *Forensic Sci. Int. Genet.* 17 (2015) 129–136.