ESCUELA DE DOCTORADO
INTERNACIONAL DE LA USC

Adrián
Ambroa Conde

Tesis doctoral

Epigenética: Desarrollo de nuevas aplicaciones en genética forense

Santiago de Compostela, 2024

Programa de doctorado en Avances y Nuevas Estrategias en Ciencias Forenses

TESIS DE DOCTORADO

EPIGENÉTICA: DESARROLLO DE NUEVAS APLICACIONES EN GENÉTICA FORENSE

Adrián Ambroa Conde

Supervisoras: Mª Victoria Lareu Huidobro y Ana Mª Freire Aradas

ESCUELA DE DOCTORADO INTERNACIONAL DE LA UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOCTORADO EN AVANCES Y NUEVAS ESTRATEGIAS EN CIENCIAS FORENSES

SANTIAGO DE COMPOSTELA

2024

Este trabajo ha sido parcialmente financiado por el proyecto de investigación BIO2013-42188-R del Ministerio de Economía, Industria y Competitividad 2014-2017, el proyecto ED481B 2018/010 de la Consellería de Cultura, Educación e Ordenación Universitaria y de la Consellería de Economía, Emprego e Industria de la Xunta de Galicia, así como por el proyecto VISAGE del programa de Investigación e Innovación Horizonte 2020 de la Unión Europea con arreglo al convenio de subvención número 740580.

AGRADECIMIENTOS

La literatura española tiene obras que han marcado la historia, siendo aún a día de hoy referentes en este arte. Si bien es cierto que este texto no será una de esas obras, me gustaría rescatar una frase de una de ellas en este documento pues, como dijo el hidalgo Don Quijote de la Mancha, "*De gente bien nacida es agradecer los beneficios que reciben, y uno de los pecados que más a dios ofende es la ingratitud*".

Una de las primeras cosas que debe agradecerse es la oportunidad concedida. La ocasión de hacer el doctorado surgió por sorpresa, gracias Maviki y Ana. A Maviki por el esfuerzo de mantener vivo el instituto en tiempos tan complicados, periodos que en la universidad parece que son eternos. Por luchar contra viento y marea contra las organizaciones que supuestamente nos amparan pero que no nos ayudan. Por el esfuerzo desinteresado para otorgar beneficios a los demás. Gracias también por tu mirada experta y crítica, siempre desde el cariño, que alimenta y cimienta las investigaciones del laboratorio. A Ana por los amplios conocimientos que ha compartido conmigo, por su amabilidad y su paciencia en la dirección de la tesis de una persona que puede ser un poco cabezona. Tu comprensión y optimismo han sido brisas cálidas que han mantenido la cometa de este doctorado siempre en alto. Por esos momentos divertidos en las comidas en horario europeo, que tan mal hemos logrado instaurar. Por esos muros invisibles, que he podido vislumbrar, en los que se han ido formando ventanas que se mantienen abiertas. Por los innumerables "donde dije digo, digo Diego" que me han enseñado que es de sabios rectificar. Ha sido un gran viaje, en cuyo transcurso he logrado adquirir alguna de las muchas virtudes que tienes y que te convierten en una grandísima investigadora.

Gracias Chris por siempre aportar un aura alegre y jovial que endulza cualquier café solo. Por los conocimientos y gran experiencia aportada y por el camino allanado que has dejado con tu esfuerzo a los nuevos que vamos llegando. Gracias Toño y Ángel que, junto a Maviki y Chris, habéis construido un lugar del que me siento tremendamente agradecido de formar parte.

Gracias a las integrantes del despacho de los listos, el *AnaPack* y María, cuyo conocimiento y buen hacer se filtra en todos los recién llegados. A la "jefecilla india" por el gran trabajo y esfuerzo que carga sobre sus espaldas, aliviando con una sonrisa el peso de los demás. Tus siempre amables y comprensivas respuestas hacen que trabajar contigo sea un camino de rosas. Aún me sigo sorprendiendo de que, con todo lo que tienes encima, siempre tengas un momento para todo el mundo, yo personalmente creo que tiene un giratiempo escondido en el bolsillo.

Gracias María, jefa de los "nanoporros", por esa energía desbordante que hace tu ritmo envidiable e inalcanzable. Esa capacidad de hacer miles de cosas en un día, cuadrando con precisión quirúrgica todas las actividades del cole y de fuera del cole, hacen que nos preguntemos cuántas horas tienen tus días. Por esos murmullos que nos acompañan un segundo después de abandonar la habitación en la que estábamos hablando y por un biberón perdido que aparece en los lugares más insospechados.

Grazas Meli pola túa destreza técnica e a paciencia que tes con estes tolos cativos que tantas crebaduras de cabeza damos. Polos teus valores morais e éticos na defensa da nosa terra. Por ser a comisión de festas de Santiago e arredores, informándonos de todos os eventos culturais da comarca. Fas máis pola cultura galega que a Xunta.

A los integrantes del cuartucho de los tontos, Jorge, Amaia, Lucía, Javi y Miguel, las minorías de nuestro laboratorio, que alegran y nutren de carcajadas todos los días laborales del año. A Jorge por su sonrisa inversa y su oído afilado para los chismes. Por ignorarnos, con razón, con sus conciertos en bucle para que el único caos de su vida sea el de su mesa. Por viajar más kilómetros que el jet privado de Taylor Swift, pero, sobre todo, por ser el compañero con el que empecé y termino este camino. Gracias por toda la ayuda brindada, espero haberte podido devolver, aunque sea una pequeña fracción de lo que has ofrecido. Gracias a Amaia por compartir su pasión y conocimientos antropológicos, por la laterización de los huesos del desván y las tomas falsas de audios que se escuchan desde la sala de PCR de investigación. Es un placer machacar huesos cochambrosos como cavernícolas con alguien que ayuda y se esfuerza tanto de forma tan desinteresada. Gracias a Lucía por llenar de sonrisas nuestros días y, siendo la máxima exponente de la ENAC (Entidad Nacional de Cartelería), de llenar nuestras paredes de carteles chulísimos. Aunque tiene un aura de calañenta, en el fondo, es un trozo de pan y una persona genial. A Javi por ser el maestro de los programas estadísticos forenses, hacedor del Euroformix y conocedor de las LRs. Por su agobio con años vista y sus capacidades esquivando los ultravioletas. Por ser el "*lore master"* oficial de las frikadas que nos apasionan. A Miguel por ser mi sustituto en la nigromancia investigadora, que siempre porta como ayuda su tijera favorita. Por su gran capacidad de aprendizaje y por el esfuerzo y las ganas de aprender mostrado, aun cuando el trabajo al inicio fue complicado. Gracias a ambos, por traer el frikismo de la fantasía a esta sala, donde todos los días se habla del Señor de los Anillos, por mucho que le pese a la pobre Lucía.

A Ángeles y Antonio por hacer cercanas y asequibles las matemáticas. Cada reunión con vosotros acababa con una sonrisa y una sensación gratificante, nuestros modelos no serían lo mismo sin vosotros. Gracias por ser tan comprensivos y dedicados, por atender nuestros desafíos y problemas con una sonrisa, palabras amables y voluntad inquebrantable. Gracias por enseñarnos tantas palabrotas matemáticas que a veces se nos hacían incomprensibles.

A las personas que han buscado otros caminos más allá de este nuestro laboratorio o me han acompañado en este viaje. A Lorena por ayudarme a dar los primeros pasos de bebé en genética forense. A Borja por su trabajo intachable y perfecto. A Vero, Iria y Vera, por sacarme un poco de casa, tarea que reconozco que es muy complicada. A las italianas, Noemi, Desiree, Giorgia, Flavia, Serena, Francesca y Emma por su alegría, por ser nuestros carabinieri de la pasta y por el intercambio de las divertidas palabras que componen nuestros idiomas. Gracias a todas las personas que han pasado por el laboratorio, ya que han aportado riqueza a mi experiencia como investigador.

Gracias a mis compañeros de carrera con los que he compartido experiencias tan memorables como los ataques de la "*cosa negra*", los ataques de risa de Iván y los días de no tanto estudio en casa de María. Con vosotros los cuatro años de Química pasaron como un suspiro en una tarde de descanso.

Gracias a mis mejores amigos, los sabios integrantes del Concilio de Kioto: Alex, Marcos, Jero, Jesús y Fredy. Por todas esas tardes, noches y madrugadas de vicio, por ser unos absolutos mancos, a excepción de Jesús, en todo a lo que jugamos. A Alex por ser tan divertido y relajado, contagias un estado de feliz tranquilidad que alivia cualquier mal. A Marcos por arrancarnos una y otra vez sonrisas con su humor ácido, por sus incontables datos extravagantes y por estar siempre presente para cualquiera que lo necesite. A Jero por ser la voz de la razón y la mesura, el punto de control y realidad de nuestras locas divagaciones. A Jesús por esos veranos pokemaníacos y esas épicas batallas en la casa del árbol donde derrotamos a incontables enemigos. Por todas las charlas sinceras y reconfortantes, por tu ayuda y apoyo incondicional. A Fredy por los cientos de horas vividas juntos en los mundos de *Azeroth* y *Tamriel*. Por las tardes y noches de series, pelis y comida insana disfrutadas. Por ser el cabronazo con el que llevo desde pequeño compartiendo una vida que no imagino sin él. Por ser mi *kyoudai*, un belicosero con un enorme corazón. En definitiva, gracias por ser esa familia que se escoge.

A mi familia por estar siempre apoyándome y ver valor en todo lo que hago. A mi madre por todo el esfuerzo realizado, aun cuando en muchas ocasiones no era bueno para su salud. Por interesarse y preguntarme por mi vida, aunque para su desgracia soy bastante parco en palabras. A mi padre por su gran corazón. Porque, aunque rosma cuando le pides algo, se preocupa porque no te falte de nada. A ambos por educarme y criarme desde el cariño, por dar sin pedir nada a cambio y por siempre cuidarme. A mi hermana, posiblemente el ser más torpe del universo, por ser una guerrera frenética. Por su fuerza de superación y energía caótica. Por ser firme en sus opiniones e intentar ayudar a quienes más quiere. Por siempre estar disponible si necesitas algo. A mis abuelos por cuidarme durante tanto tiempo, por preocuparse y por todos esos caprichos ocultos a los ojos de mis padres en una pequeña palma. Es posible que vaya un poco a mi bola, pero eso no cambia lo mucho que os quiero.

A mi compañera de vida, Alejandra. Durante mucho tiempo pensé que el amor solo eran las reacciones químicas controladas por las hormonas que había dado durante la carrera, pero me mostraste lo equivocado que estaba. Gracias a ti esa palabra ha ganado múltiples significados: anhelar una vida juntos, imaginar un futuro a tu lado, que inundes mis pensamientos nada más despertarme y antes de acostarme, querer verte y estar contigo todos los días, sonreír o que se me curen todos los males nada más verte. Te has convertido en mi fuerza, en la catalizadora de mi motivación, tus ojos se han convertido en mi hogar. Gracias a tu inteligencia y mente afilada mi mundo ha ganado riqueza intelectual, reactivando en mí una característica desactivada desde la infancia, la curiosidad por descubrir y aprender cosas nuevas. Siempre he sido una persona casera, siendo generoso conmigo mismo, pero contigo ha surgido la motivación y las ganas de descubrir y conocer cosas nuevas. Mi vida ha ganado una alegría dulce como la miel, que hace que todo a mi alrededor se pare para permitirme saborear con calma cada momento contigo. Como le dijo Arwen a Aragorn "*Prefiero vivir una vida mortal a tu lado, que enfrentarme a todas las Edades de este mundo solo*". Gracias por apoyarme y estar a mi lado, te quiero.

A veces las palabras, seas o no diestro con ellas, no logran transmitir todo lo que sentimos. En mi caso quedan muchas cosas por decir, pero esos sentimientos me los guardo para mí. Gracias a todos por todos esos momentos, sonrisas, conocimientos y experiencias vividas que atesoraré con cariño.

Vida antes que muerte, fuerza antes que debilidad, viaje antes que destino.

> “El antiguo código de los Caballeros Radiantes reza: Viaje antes que destino. Algunos lo consideran un simple lugar común, pero es mucho más. Un viaje incluirá dolor y fracaso. No son solo los pasos adelante los que debemos aceptar, sino también los traspiés. Las dificultades. El conocimiento de que fracasaremos. De que haremos daño a quienes nos rodean. Pero si nos detenemos, si aceptamos la persona que somos al caer, el viaje concluye. Ese fracaso pasa a ser nuestro destino. Amar el viaje implica no aceptar ese final. He descubierto, por medio de dolorosas experiencias, que el paso más importante que puede dar alguien es siempre el siguiente.”
>
> Juramentada de Brandon Sanderson

ÍNDICE

LISTA DE PUBLICACIONES

Artículo 1: A common epigenetic clock from childhood to old age

A. Freire-Aradas[1], L. Girón Santamaría[2], A. Mosquera-Miguel[2], A. Ambroa-Conde[2], C. Phillips[2], M. Casares de Cal[3], A. Gómez-Tato[3], J. Álvarez-Dios[3], E. Pospiech[4], A. Aliferi[5], D. Syndercombe Court[5], W. Branicki[6], M. V. Lareu[2].

Afiliación:

[1] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain. Electronic address: ana.freire@usc.es.
[2] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain.
[3] Faculty of Mathematics, University of Santiago de Compostela, Spain.
[4] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland.
[5] King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom.
[6] Laboratory of Anthropology, Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland.

Forensic Science International: Genetics, volumen 60, página 102743, 2022, PMID: 35777225, DOI: 10.1016/j.fsigen.2022.102743

Referencia completa:

Freire-Aradas A, Girón-Santamaría L, Mosquera-Miguel A, Ambroa-Conde A, Phillips C, Casares de Cal M, Gómez-Tato A, Álvarez-Dios J, Pospiech E, Aliferi A, Syndercombe Court D, Branicki W, Lareu MV. A common epigenetic clock from childhood to old age. Forensic Sci Int Genet. 2022 Sep;60:102743. doi: 10.1016/j.fsigen.2022.102743. Epub 2022 Jun 25. PMID: 35777225.

Contribución específica en la publicación:

Selección de marcadores epigenéticos, procesamiento y análisis de datos, evaluación estadística y revisión del manuscrito original.

Índices de calidad:

La revista donde fue publicada el artículo 1 presenta un índice de impacto de 3,1 (2022 Journal Impact Factor), un índice CiteScore de 7,9 en 2022 (calculado por Scopus el 05 de mayo de 2023) y las siguientes categorías: cuartil 1 (Q1) en *Genetics* y *Pathology and Forensic Medicine* (SJR 2022 1,39) calculado por Scimago: https://www.scimagojr.com/journalsearch.php?q=26950&tip=s

Autorización de la revista:

La revista *Forensic Science International: Genetics*, perteneciente a la editorial Elsevier, donde se ha publicado el artículo 1, permite la reutilización del artículo por parte del autor como parte de su tesis: https://www.elsevier.com/about/policies-and-standards/copyright

Artículo 2: Epigenetic age estimation in saliva and in buccal cells

A. Ambroa-Conde[1], L. Girón Santamaría[1], A. Mosquera-Miguel[1], C. Phillips[1], M. A. Casares de Cal[2], A. Gómez-Tato[2], J. Álvarez-Dios[3], M. de la Puente[1], J. Ruiz-Ramírez[1], M. V. Lareu[1], A. Freire-Aradas[4].

Afiliación:

[1] Unidad de Genética Forense, Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, España.
[2] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain.
[3] Faculty of Mathematics, University of Santiago de Compostela, Spain.
[4] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain. Electronic address: ana.freire@usc.es.

Forensic Science International: Genetics, volumen 61, página 102770, 2022, PMID: 36057238, DOI: 10.1016/j.fsigen.2022.102770

Referencia completa:

Ambroa-Conde A, Girón-Santamaría L, Mosquera-Miguel A, Phillips C, Casares de Cal MA, Gómez-Tato A, Álvarez-Dios J, de la Puente M, Ruiz-Ramírez J, Lareu MV, Freire-Aradas A. Epigenetic age estimation in saliva and in buccal cells. Forensic Sci Int Genet. 2022 Nov;61:102770. doi: 10.1016/j.fsigen.2022.102770. Epub 2022 Aug 27. PMID: 36057238.

Contribución específica en la publicación:

Recolección de muestras biológicas, selección de marcadores epigenéticos, diseño y desarrollo experimental, procesamiento y análisis de datos, evaluación estadística y redacción del manuscrito original.

Índices de calidad:

La revista donde fue publicada el artículo 1 presenta un índice de impacto de 3,1 (2022 Journal Impact Factor), un índice CiteScore de 7,9 en 2022 (calculado por Scopus el 05 de mayo de 2023) y las siguientes categorías: cuartil 1 (Q1) en *Genetics* y *Pathology and Forensic Medicine* (SJR 2022 1,39) calculado por Scimago: https://www.scimagojr.com/journalsearch.php?q=26950&tip=s

Autorización de la revista:

La revista *Forensic Science International: Genetics*, perteneciente a la editorial Elsevier, donde se ha publicado el artículo 2, permite la reutilización del artículo por parte del autor como parte de su tesis: https://www.elsevier.com/about/policies-and-standards/copyright

Artículo 3: Development of an epigenetic age predictor for costal cartilage with a simultaneous somatic tissue differentiation system

A. Freire-Aradas[1], M. Tomsia[2], D. Piniewska-Róg[3], A. Ambroa-Conde[4], M. A. Casares de Cal[5], A. Pisarek[6], A. Gómez-Tato[5], J. Álvarez-Dios[7], E. Pośpiech[8], W. Parson[9], M. Kayser[10], C. Phillips[4], W. Branicki[11].

Afiliación:

[1]Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain. Electronic address: ana.freire@usc.es.
[2]Department of Forensic Medicine and Forensic Toxicology, Medical University of Silesia, Katowice, Poland.
[3]Department of Forensic Medicine, Jagiellonian University Medical College, Kraków, Poland.
[4]Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain.
[5]CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain.
[6]Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland.
[7]Faculty of Mathematics, University of Santiago de Compostela, Spain.
[8]Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland; Department of Forensic Genetics, Pomeranian Medical University in Szczecin, Poland.
[9]Institute of Legal Medicine, Medical University of Innsbruck, Austria; Forensic Science Program, Pennsylvania State University, PA, USA.
[10]Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands.
[11]Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland; Institute of Forensic Research, Kraków, Poland. Electronic address: wojciech.branicki@uj.edu.pl.

Forensic Science International: Genetics, volumen 67, página 102936, 2023 DOI: 10.1016/j.fsigen.2023.102936

Referencia completa:

Freire-Aradas A, Tomsia M, Piniewska-Róg D, Ambroa-Conde A, Casares de Cal MA, Pisarek A, Gómez-Tato A, Álvarez-Dios J, Pośpiech E, Parson W, Kayser M, Phillips C, Branicki W. Development of an epigenetic age predictor for costal cartilage with a simultaneous somatic tissue differentiation system. Forensic Sci Int Genet. 2023 Nov;67:102936. doi: 10.1016/j.fsigen.2023.102936. Epub 2023 Sep 29. PMID: 37783021.

Contribución específica en la publicación:

Procesamiento y análisis de datos, evaluación estadística, revisión del manuscrito original.

Índices de calidad:

La revista donde fue publicada el artículo 3 presenta actualmente un índice de impacto de 3,2 (2023 Journal Impact Factor), un índice CiteScore de 7,5 (calculado por Scopus el 17 de junio de 2024) y las siguientes categorías: cuartil 1 (Q1) en *Genetics* y *Pathology and Forensic Medicine* (SJR 2023 1,34) calculado por Scimago: https://www.scimagojr.com/journalsearch.php?q=26950&tip=s

Autorización de la revista:

La revista *Forensic Science International: Genetics*, perteneciente a la editorial Elsevier, donde se ha publicado el artículo 3, permite la reutilización del artículo por parte del autor como parte de su tesis: https://www.elsevier.com/about/policies-and-standards/copyright

Artículo 4: Inference of tobacco and alcohol consumption habits from DNA methylation analysis of blood

A. Ambroa-Conde[1], M. A. Casares de Cal[2], A. Gómez-Tato[2], O. Robinson[3], A. Mosquera-Miguel[1], M. de la Puente[1], J. Ruiz-Ramírez[1], C. Phillips[1], M. V. Lareu[1], A. Freire-Aradas[4].

Afiliación:

[1]Forensic Genetics Unit, Institute of Forensic Sciences, Universidade de Santiago de Compostela, Spain.
[2]CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain.
[3]MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK.
[4]Forensic Genetics Unit, Institute of Forensic Sciences, Universidade de Santiago de Compostela, Spain. Electronic address: ana.freire@usc.es.

Forensic Science International: Genetics, volumen 70, página 103022, 2024 DOI: 10.1016/j.fsigen.2024.103022

Referencia completa:

Ambroa-Conde A, Casares de Cal MA, Gómez-Tato A, Robinson O, Mosquera-Miguel A, de la Puente M, Ruiz-Ramírez J, Phillips C, Lareu MV, Freire-Aradas A. Inference of tobacco and alcohol consumption habits from DNA methylation analysis of blood. Forensic Sci Int Genet. 2024 May;70:103022. doi: 10.1016/j.fsigen.2024.103022. Epub 2024 Jan 28. PMID: 38309257.

Contribución específica en la publicación:

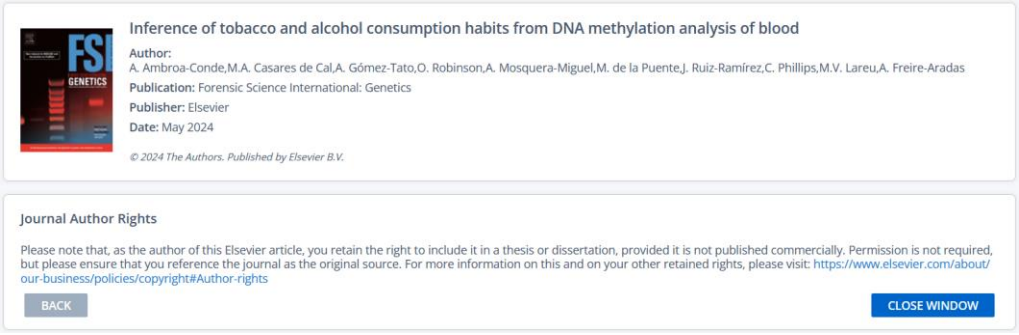
Gestión de datos de metilación de ADN de la base de datos empleada, selección de marcadores epigenéticos, procesamiento y análisis de datos, evaluación estadística y redacción del manuscrito original.

Índices de calidad:

La revista donde fue publicada el artículo 4 presenta actualmente un índice de impacto de 3,2 (2023 Journal Impact Factor), un índice CiteScore de 7,5 (calculado por Scopus el 17 de junio de 2024) y las siguientes categorías: cuartil 1 (Q1) en *Genetics* y *Pathology and Forensic Medicine* (SJR 2023 1,34) calculado por Scimago: https://www.scimagojr.com/journalsearch.php?q=26950&tip=s

Autorización de la revista:

La revista *Forensic Science International: Genetics*, perteneciente a la editorial Elsevier, donde se ha publicado el artículo 4, permite la reutilización del artículo por parte del autor como parte de su tesis: https://www.elsevier.com/about/policies-and-standards/copyright



ABREVIATURAS

%CP±PI: Porcentaje de clasificaciones correctas dentro del intervalo de predicción.

5-caC: 5-carboxilcitosina.

5-fC: 5-formilcitosina.

5-hmC: 5-hidroxilmetilcitosina.

5-mC: 5-metilcitosina.

a.C.: antes de Cristo.

ADN: Ácido desoxirribonucleico.

ADNmt: Ácido desoxirribonucleico mitocondrial.

AGE: Producto final de glicación avanzada, del inglés *Advanced Glycation End products*.

AIM: Marcadores informativos de ancestralidad, del inglés *Ancestry-Informative-Markers*.

APS: persulfato amónico.

ARN: Ácido ribonucleico.

ARNnc: Ácido ribonucleico no codificante.

ATP: Adenosín trifosfato.

AUC: Área bajo la curva, del inglés *Area Under the Curve*.

BER: Ruta de reparación por escisión de bases, del inglés *Base Excision Reparation*.

BGA: Origen biogeográfico, del inglés *Biogeographic Ancestry*.

BMI: Índice de masa corporal, del inglés *Body Mass Index*.

Bp: par de bases, del inglés *base pair*.

BR: Rango amplio, del inglés *Broad Range*.

CEPH: *Centre d'Etude du Polymorphisme Humain*.

CGI: islas CpG.

CO_2: dióxido de carbono.

CpG: Dinucleótidos citosina y guanina enlazados por fosfato.

ddNTP: Didesoxinucleótidos.

ddPCR: Reacción en cadena de la polimerasa digital, del inglés *droplet digital PCR*.

dGTP: Deoxiguanosina trifosfato.

DNMT1: Ácido desoxirribonucleico metiltransferasa 1.

DNMT3a: Ácido desoxirribonucleico metiltransferasa 3 alfa.

DNMT3b: Ácido desoxirribonucleico metiltransferasa 3 beta.

EVC: Características visibles externas, del inglés *External Visible Characteristics.*

FDP: Fenotipado forense de ADN, del inglés *Forensic DNA Phenotyping*.

GWAS: Estudio de asociación de genoma completo, del inglés *Genome-Wide Association Study*.

HGDP: Proyecto de Diversidad del Genoma Humano, del inglés *Human Genome Diversity Project*.

HGP: Proyecto Genoma Humano, del inglés *Human Genome Project.*

HS: Alta sensibilidad, del inglés *High Sensitivity*.

HV1: Región hipervariable 1 del ADN mitocondrial.

HV2: Región hipervariable 2 del ADN mitocondrial.

INE: Instituto Nacional de Estadística.

Kb: Kilobase, hace referencia a 1000 pares de bases de ADN o de ARN.

m/z: Relación masa-carga.

MAE_{media}: Error medio absoluto, del inglés *Mean Absolute Error*.

$MAE_{mediana}$: Error mediano absoluto, del inglés *Median Absolute Error*.

MALDI: Desorción/ionización láser asistida por matriz, del inglés *Matrix-Assisted Laser Desorption/Ionization*.

miRNA: Micro ácido ribonucleico.

MHs: Microhaplotipos.

miARN: Micro ácido ribonucleico.

MLPs: Sondas multi locus, del inglés *Multi Locus Probes*.

MPS: Secuenciación masiva en paralelo, del inglés *Massive Parallel Sequencing*.

mRNA: Ácido ribonucleico mensajero.

MZ: Monocigótico.

NDNAD: Base de datos nacional de Reino Unido, del inglés *UK National DNA Database*.

Ng: Nanogramos.

OEDE: Observatorio Español de las Drogas y las Adicciones.

PCA: Análisis de las Componentes Principales.

PCR: Reacción en Cadena de la Polimerasa, del inglés *Polymerase Chain Reaction.*

Pg: Picogramos.

PPi: Pirofosfato.

qPCR: Reacción en cadena de la polimerasa cuantitativa, del inglés *quantitative Polymerase Chain Reaction*.

RFLP: Polimorfismos de longitud de fragmentos de restricción, del inglés *Restriction Fragment Length Polymorphism*.

RFUs: Unidades de fluorescencia relativas, del inglés *Relative Fluorescence Units*.

RMSE: Error cuadrático medio, del inglés *Root-Mean-Squared Error*.

SBE: Extensión de un único nucleótido, del inglés *Single Base Extension*.

sjTREC: Círculos de escisión del receptor de células T, del inglés *Signal-joint T-cell Receptor Excision Circles*.

SLP: Sondas unilocus, del inglés *Single Locus Probes*.

SNP: Polimorfismos de un solo nucleótido, del inglés *Single Nucleotide Polymorphism.*

STR: *Short Tandem Repeats*.

TDG: Timina ADN glicosilasa, del inglés *Thymine DNA Glycosylase*.

TET: *Ten-Eleven Translocation.*

TOF: Tiempo de vuelo, del inglés *Time Of Flight*.

VISAGE: *Visible Attributes Through Genomics*.

VNTR: Repeticiones en tándem de número variable, del inglés *Variable Number of Tandem Repeats*.

WGBS: Secuenciación con bisulfito de genoma completo, del inglés *Whole Genome Bisulfite Sequencing.*

OMS: Organización Mundial de la Salud.

X-STR: *Short Tandem Repeats* del cromosoma X.

YHRD: *Y-Chromosome Haplotype Reference Database*.

Y-SNP: Polimorfismos de un solo nucleótido del cromosoma Y.

Y-STR: *Short Tandem Repeats* del cromosoma Y.

RESUMEN

RESUMEN

Dentro de las Ciencias Forenses la genética ha adquirido con los años una gran importancia a nivel judicial. La solicitud de pruebas de ADN ha inundado los juzgados de todo el mundo a fin de proporcionar una respuesta a relaciones de parentesco biológico, resolución de casos criminales e identificación humana. El nacimiento y el auge de esta disciplina se cimientan en diversos descubrimientos de gran repercusión en el ámbito de la biología molecular. Los patrones de herencia genética, el descubrimiento de la molécula de ADN y su estructura, la identificación de polimorfismos genéticos y el desarrollo de la reacción en cadena de la polimerasa han definido un ecosistema de conocimientos que han llevado a la genética forense a una era dorada. Pero estos solo eran los primeros pasos hacia la cima, el avance no hizo más que acelerarse alcanzando nuevas cotas gracias al esfuerzo de la comunidad científica. En menos de 100 años, los estudios de parentesco e identificación pasaron de emplear el grupo sanguíneo AB0, dentro del contexto de la hemogenética, al descubrimiento e implementación de polimorfismos genéticos, como son los *Short Tandem Repeats* (STRs), *Single Nucleotide Polymorphism* (SNPs), microhaplotipos, dando paso a la genética forense. Cada uno de estos polimorfismos presentan ventajas y desventajas que han ido, a lo largo del tiempo, configurando sus aplicaciones en el campo. La realidad de este ámbito es que todas las aplicaciones desarrolladas deben ajustarse a unas condiciones específicas, siendo de gran importancia su evaluación a la hora de decidir qué tipo de polimorfismos estudiar en cada caso y sobre qué material genético analizarlos. Con estas características en mente se han ido definiendo aplicaciones que recurren al estudio de estos marcadores tanto en el ADN autosómico, como en los cromosomas sexuales e incluso en el ADN extra nuclear, como es el ADN mitocondrial. A nivel identificativo, cuando se dispone de un perfil de referencia para comparar con el perfil genético obtenido en una muestra biológica, los STRs abarcan el protagonismo de los análisis. Pero existen situaciones en las que no es posible llevar a cabo esta comparación, requiriéndose de información adicional para reducir el número de sospechosos. Por tanto, se han desarrollado técnicas cuyo objetivo es el fenotipado forense, aportando información de características como, la inferencia del origen biogeográfico, la predicción de características físicas o la estimación de la edad, así como estilos de vida del donante. Gracias al desarrollo de estas herramientas, el estudio de la información contenida en nuestro ADN nos permite inferir en la actualidad la procedencia genética de un individuo a nivel intercontinental (África Subsahariana, Este de Asia, América, Oceanía o Europa), así como su color de ojos, pelo y piel. Además de los marcadores genéticos que permiten predecir estos rasgos, también se han comenzado a explorar marcadores epigenéticos, como es la metilación del ADN, teniendo también como finalidad el fenotipado forense. Este biomarcador ha sido ampliamente estudiado en los últimos años con el objetivo de aportar información que el genoma por sí mismo no puede, definiéndose el epigenoma.

El epigenoma ha destacado por sus efectos sobre la regulación genética, ampliando nuestro entendimiento en relación con nuestro genoma y la transmisión de su información. Este código de regulación está compuesto por diferentes marcas epigenéticas siendo la más relevante en el ámbito forense, la metilación del ADN. La metilación del ADN es, como su propio nombre

indica, la introducción de un grupo metilo en una de las bases nitrogenadas que componen la secuencia de nuestro genoma, en concreto, en el carbono 5' de aquellas citosinas seguidas de guaninas, posiciones denominadas como CpGs. La metilación del ADN es una marca dinámica que está influenciada por el genoma, el ambiente y factores estocásticos, lo que le confiere ciertas características muy interesantes tanto a nivel clínico como forense. Al ser una modificación que se encuentra "sobre" el ADN, los análisis convencionales no pueden detectarlo, por lo que para su estudio se requiere, de forma general, de un pretratamiento que permita diferenciar entre las citosinas metiladas y no metiladas. Una de estas metodologías es la conversión con bisulfito sódico que genera una modificación detectable en la secuencia del ADN. Con este pretratamiento se convierten las citosinas no metiladas en uracilos mediante una desaminación, proceso que se ve bloqueado si la citosina está metilada, obteniendo tras la amplificación, timinas en lugar de citosinas no metiladas y citosinas donde existe metilación. Gracias a esta transformación se han podido adaptar metodologías como la pirosecuenciación, EpiTYPER, minisecuenciación y la secuenciación masiva en paralelo (MPS) para el análisis de la metilación del ADN. Centrándonos en el campo que nos compete, el interés en el estudio de esta marca epigenética radica en su correlación con ciertos factores que podrían emplearse para generar información sobre una muestra biológica o sobre el donante de dicho vestigio. Con esto en mente, se han definido diversas aplicaciones entre las que encontramos la discriminación de gemelos monocigóticos, la identificación de tejidos, la estimación de la edad cronológica y la inferencia de estilos de vida. Estas aplicaciones, unas más que otras, han sido ampliamente estudias, estando algunas de ellas a las puertas de su implementación en casuística.

Respecto a la discriminación de gemelos monocigóticos, todavía no se ha logrado identificar posiciones comunes que muestren diferencias significativas, planteándose más como un proceso específico que debería realizarse para cada par de gemelos. Una situación diferente se presenta para la identificación de tejidos. Una de las características de la metilación del ADN es su especificidad de tejido, mostrando las distintas poblaciones celulares patrones de metilación diferentes. Esto ha propiciado el estudio de este biomarcador para evaluar el tejido de origen de una muestra biológica como una alternativa para las metodologías actualmente estandarizadas que requieren de la destrucción de parte de la muestra o que solo permiten identificar un limitado número de tejidos concretos. El empleo de marcadores que presentan patrones de metilación muy diferenciados entre tejidos ha permitido discriminar entre semen, sangre, saliva, fluido vaginal y sangre menstrual, ampliando el número de tejidos identificables de interés forense. Por otro lado, gracias a la relación observada entre patrones de metilación y envejecimiento se han identificado marcadores correlacionados con la edad cronológica, abriendo las puertas a la que hoy en día es la aplicación más estudiada de la metilación del ADN, la estimación de la edad. Su estudio ha derivado en la generación de múltiples modelos de predicción, denominados relojes epigenéticos, basados en muestras de diferentes tejidos, obteniéndose, en la actualidad, para los diferentes vestigios biológicos analizados errores cercanos a los ± 3 años. Por último, una de las aplicaciones más novedosas de este biomarcador es la inferencia de estilos de vida, centrándose principalmente, por el momento, en el estudio del consumo de tabaco y alcohol. El efecto general que produce el consumo de estas sustancias

sobre el epigenoma y la reversibilidad de los patrones de metilación de ciertas posiciones asociadas a este consumo presentan un desafío en el desarrollo de esta aplicación. A pesar de ello, se ha comenzado a trabajar en el desarrollo de modelos de clasificación para estos hábitos.

La búsqueda de marcadores y la generación de modelos de predicción ha proporcionado conocimiento alrededor de la metilación del ADN más allá de sus objetivos principales. Gracias a todos los trabajos desarrollados se han definido ciertas condiciones de gran importancia que deben tenerse en cuenta a la hora de estudiar este biomarcador, pero también ciertas limitaciones que aún deben ser abordadas. Una de ellas ha sido las diferencias observadas en los niveles de metilación cuando éstos son evaluados con plataformas diferentes. Estas discrepancias han propiciado la búsqueda de métodos que permitan ajustar los valores de metilación obtenidos con un equipo para ser empleados en modelos de predicción que han sido desarrollados en base a otro diferente. Además, el tipo de tejido analizado es crucial en los estudios de metilación del ADN, destacando sobre todo la composición celular de las muestras, característica que puede condicionar el éxito del análisis. Esta condición es de gran importancia en el ámbito forense siendo necesario ampliar el conjunto de tejidos analizados a fin de abarcar la realidad del campo para poder afrontar con mayor seguridad el análisis de la metilación de muestras desconocidas. Por tanto, aplicaciones como la estimación de la edad o inferencia de estilos de vida, son dependiente de tejido, no siendo posible aplicar modelos de predicción desarrollados en base a un tejido concreto a otros, debido a las diferencias en los patrones de metilación observadas. Además, esto puede condicionar el análisis de muestras que presentan mezclas de tejidos, por lo que se deben abordar sistemas que aseguren una evaluación correcta de muestras con proporciones celulares desconocidas. Por otro lado, los estudios de estimación de la edad se han centrado en ciertos tejidos, siendo necesaria la generación de modelos de predicción para el mayor número de vestigios forenses posibles. Por último, aunque es cierto que empiezan a generarse los primeros modelos de inferencia de estilos de vida, debe tenerse en cuenta que la mayoría de ellos presentan un sesgo asociado al consumo. Por el momento la mayoría de los modelos construidos se centran en la clasificación de los grupos extremos de las categorías establecidas, obviando o teniendo dificultades para la correcta clasificación de las categorías intermedias, no representando la totalidad de la población en dichos modelos. Por tanto, deben buscarse alternativas que permitan predecir correctamente a los grupos de mayor interés a nivel forense, los más individualizantes, teniendo en cuenta la variedad de consumo existente en la población.

Con esto en mente, esta tesis doctoral se centra en el estudio de la metilación del ADN, siendo su objetivo general la búsqueda y selección de marcadores correlacionados con distintas aplicaciones forenses de este biomarcador a fin de construir modelos de estimación de la edad, identificación de tejidos e inferencia de estilos de vida. Como resultado del trabajo desarrollado se presentan cuatro artículos científicos en los que se engloban tres modelos de predicción de la edad cronológica, dos modelos de identificación de tejido y dos modelos de inferencia de estilos de vida. A su vez, teniendo en cuenta las discrepancias observadas en los valores de metilación obtenidos con metodologías semicuantitativas cuando éstas son evaluadas

empleando equipos diferentes, se ha construido un modelo estadístico de transformación con el objetivo de extender el uso de uno de los modelos desarrollados.

En el contexto de la estimación de la edad, la sangre ha sido el tejido de preferencia para el desarrollo de modelos de predicción monopolizando durante mucho tiempo la actualidad del campo. Además, existen otros puntos críticos a tener en cuenta a la hora de desarrollar relojes epigenéticos, como por ejemplo el rango de edad analizado. Muchos de los modelos desarrollados presentan amplios rangos de edades, pero, de forma consistente, los menores de edad representan un bajo porcentaje del número total de individuos. Esta infrarrepresentación deriva en una inferencia de las tendencias en los patrones de metilación en dichas edades. Esta inferencia o extrapolación puede resultar errónea en aquellos casos en los que se trabaje con marcadores con patrones de metilación que difieren entre menores y adultos; cómo, por ejemplo, cambios exponenciales de la metilación con la edad en menores, pero que siguen un ritmo de evolución más estable en adultos. Por tanto, en el **artículo 1** se presenta un modelo de regresión cuantil para la estimación de la edad en base a muestras de sangre para un rango de 2 a 104 años representado por alrededor de 10 individuos por edad. Dicho modelo está compuesto por siete posiciones CpG (*ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* y cg07082267) y presenta errores de predicción de entre ± 3,36 años y ± 3,75 años para los diferentes acercamientos matemáticos planteados.

Más allá de la sangre, se han generado modelos de predicción de la edad con otros tejidos de interés forense, entre los que encontramos la saliva y la mucosa oral. Estos tejidos presentan una complejidad intrínseca, su composición celular, compuesta por proporciones variables de células sanguíneas y células epiteliales. Esta diferencia observada entre muestras conlleva una falta de uniformidad entre los patrones de metilación de individuos de las mismas edades. La mayor parte de los modelos desarrollados se centra específicamente en uno de estos tejidos, saliva o mucosa oral, lo que podría no representar correctamente la variabilidad de dichos tejidos a nivel celular y los haría ineficientes ante muestras mezcla de ambos tejidos, como son las colillas de cigarrillo. Ante esta situación, en el **artículo 2**, se presentan dos acercamientos diferentes ante el tratamiento de muestras de la cavidad oral en genética forense para el estudio de la metilación del ADN. La disponibilidad de modelos independientes para dichos tejidos se beneficiaría de un modelo de identificación que permitiese conocer si el vestigio analizado procede de saliva o de hisopo bucal (mucosa oral), a fin de aplicar el modelo de predicción de la edad más adecuado. Por tanto, se desarrolló un modelo de regresión logística compuesto por dos posiciones CpG (*HUNK* y *RUNX1*) para la identificación de tejido que presenta un porcentaje de clasificaciones correctas del 88,59%. Pese al alto porcentaje de clasificación no se asegura una mejora de la predicción de la edad de la muestra, por lo que se propone otro acercamiento generando un modelo combinando los tejidos analizados. Por tanto, se construyó un modelo de regresión cuantil para la predicción de la edad para muestras de saliva e hisopo bucal compuesto por siete posiciones CpG (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* y *EDARADD*). Dicho modelo presenta un error mediano absoluto de ± 3,54 años, en consonancia con los resultados observados en el campo (aproximadamente ± 3 años). La introducción de la información del tejido de origen como variable del modelo no produjo una

mejoría en los errores obtenidos, descartándose como variable informativa para este modelo. Teniendo en cuenta que el tejido es una variable general, la atención debería centrarse en la gran repercusión que tiene la composición celular en los patrones de metilación detectados, siendo conveniente evaluar esta última como variable de los modelos de predicción basados en metilación del ADN. Además, en vista de la discontinuidad de soporte para el secuenciador capilar empleado para el desarrollo de este modelo (ABI3130xl) y teniendo en cuenta las discrepancias en los niveles de metilación observados entre plataformas (descritas en la bibliografía, así como en el presente estudio), se generó un método de transformación que permitiese extender su uso a un secuenciador más actual, el ABI3500. Las diferencias observadas no estaban sólo asociadas al equipo, por lo que fue necesario transformar los niveles de metilación de forma específica para cada marcador. Los modelos desarrollados, recogidos en el **Anexo I**, permiten la correcta predicción de la edad de muestras analizadas con el ABI3500 empleando el modelo de saliva e hisopo bucal original, extendiendo así la aplicabilidad de dicho modelo.

Por último, en lo referente a estimación de la edad, se han comenzado a explorar otros tejidos de interés forense, desarrollándose en los últimos años modelos en base a hueso, diente, pelo y uñas. En el **artículo 3** se presenta el primer modelo de predicción de la edad basado en muestras de cartílago. El modelo de regresión cuantil desarrollado se compone de tres posiciones CpG (*FHL2*, *TRIM59* y *KLF14*) y presenta un error de ± 4,41 años. Además, teniendo en cuenta la importancia del tejido de origen, se desarrolló un modelo de identificación con un 98,72% de clasificaciones correctas para muestras de sangre, hisopo bucal, hueso y cartílago empleando ocho posiciones CpG (*EDARADD*, *TRIM59*, *ELOVL2*, *MIR29B2CHG*, *PDE4C*, *ASPA*, *FHL2* y *KLF14*). Este modelo de identificación está compuesto por posiciones CpG contenidas en genes comúnmente correlacionados con la edad, abriendo las puertas a un análisis conjunto de marcadores de tejido y edad gracias al uso de plataformas de secuenciación masiva en paralelo.

El amplio número de estudios y los buenos resultados obtenidos en la estimación de la edad e identificación de tejidos han abierto las puertas a la evaluación de otros factores asociados al epigenoma, como el estudio de los estilos de vida. Teniendo en cuenta la situación del campo con respecto a la inferencia del consumo de tabaco y alcohol, en el **artículo 4** se han desarrollado modelos de clasificación de estado de fumador y bebedor. En primer lugar, se evaluó la construcción de modelos logísticos multinomiales que permitiesen clasificar simultáneamente las categorías establecidas (tabaco: no fumador, exfumador y fumador; alcohol: no bebedor, bebedor moderado y bebedor alto). Los resultados obtenidos mostraron una clasificación errónea de los grupos intermedios de ambos estilos de vida, presentando una cierta tendencia a ser clasificados como no consumidores. En base a esto se construyeron modelos logísticos binomiales agrupando la categoría intermedia y los no consumidores en un mismo grupo. En el caso de consumo de tabaco se desarrolló un modelo de clasificación compuesto por dos posiciones CpG (*AHRR* y cg01940273) que presenta un porcentaje de clasificaciones correctas de 86,49%. Por otro lado, en relación con el consumo de alcohol, se desarrolló un modelo de clasificación compuesto por tres posiciones CpG (*SLC7A11*,

cg0886875 y *MIR4435–2HG*) que presenta un porcentaje de clasificaciones correctas del 74,26%. Por último, teniendo en cuenta el efecto general que producen estos estilos de vida en el epigenoma y su relación con un envejecimiento acelerado, se evaluó si estas condiciones de consumo podrían afectar a los errores de predicción de la edad. Para ello, se construyó un modelo de regresión cuantil de predicción de la edad en base a cinco posiciones CpG previamente reportadas y se evaluaron las diferencias observadas en los residuos generados para las categorías de consumo analizadas, no observándose diferencias significativas.

Gracias al trabajo desarrollado durante esta tesis doctoral se han logrado construir modelos de predicción a partir del ADN de interés forense para la mayoría de las aplicaciones asociadas a la metilación del ADN en relación con el campo. En primer lugar, se ha conseguido identificar marcadores correlacionados con la edad en varios tejidos, marcadores tejido-específicos y posiciones asociadas al consumo de tabaco y alcohol. Esto ha permitido el desarrollo de diversos modelos de interés forense que intentan abordar, en mayor o menor medida, ciertas limitaciones del campo. Los modelos de predicción de la edad construidos presentan errores cercanos a ± 3-4 años, resultados en consonancia con los observados actualmente en el campo. A fin de mejorar los resultados existentes, se ha desarrollado un modelo en base a muestras de sangre compuesto por un amplio rango de edades, en las que cada edad está representada por alrededor de diez individuos. Teniendo en cuenta la complejidad de los tejidos procedentes de la cavidad oral, se ha construido un modelo que combina dos vestigios biológicos, saliva y mucosa oral. La aplicabilidad y vida útil de dicho modelo fue extendida mediante la transformación de los niveles de metilación, permitiendo la correcta predicción de la edad de muestras analizadas con un secuenciador capilar diferente del empleado en la construcción del modelo de predicción. Finalmente, en lo referente a estimación de la edad, se ha abordado el estudio de un tejido de interés forense que aún no había sido analizado, construyéndose un modelo de predicción de la edad en base a muestras de cartílago. No obstante, el origen de un resto biológico analizado es de gran importancia en el estudio de la metilación del ADN, por tanto, se han desarrollado modelos logísticos de identificación de tejido. Por un lado, se ha construido un modelo que permite clasificar entre hisopo bucal y saliva y, por otra parte, un modelo de identificación para sangre, hisopo bucal, hueso y cartílago, aportando resultados sobre el estudio de tejidos complejos o la identificación de nuevos tejidos. Por último, se ha abordado la inferencia de estilos de vida, desarrollándose modelos de clasificación para el consumo de tabaco y alcohol que buscan representar las diferentes categorías observadas en la población general. Gracias a todos los trabajos que componen esta tesis doctoral se han evaluado o definido nuevos aspectos del estudio de la metilación del ADN en genética forense, colaborando en la mejora o avance del estudio de este biomarcador.

INTRODUCCIÓN

1. INTRODUCCIÓN

1.1 Historia de la genética forense

Los actos delictivos son sucesos que han acompañado a la humanidad desde su origen. En 2015, tras el análisis de unos restos cadavéricos datados del periodo Chibaniano, también conocido como Pleistoceno Medio, se clasificó la muerte de un homínido que vivió hace 430.000 años, debido a las fracturas que presentaba la reconstrucción de su cráneo, como un asesinato (1). Si bien es cierto que las circunstancias de este suceso son desconocidas, siendo su clasificación como "asesinato" más una hipótesis que una certeza, se puede afirmar de forma indiscutible que la violencia ha estado muy presente en nuestra historia. Pero estos actos delictivos han ido acompañados de procesos que impartían castigos, no teniendo que ser todos ellos justos. Conocer un poco la historia de un ámbito nos proporciona una visión más amplia del mismo, ofreciendo un relato del camino seguido hasta llegar al punto en el que nos encontramos. Por tanto, descubramos brevemente el camino recorrido por las ciencias forenses a lo largo de los siglos. Uno podría pensar que no tenemos por qué remontarnos mucho en el pasado para comentar los inicios de esta disciplina, pero quizás nos sorprenda la antigüedad que tienen muchas de las cosas que hoy consideramos cotidianas en este campo. Es interesante destacar que los primeros registros de procesos judiciales se remontan al antiguo Egipto y datan del 3.000 a. C. Dichos procesos se llevaban a cabo por tribunales conformados por sacerdotes y funcionarios que se basaban en la legislación existente para impartir justicia (2). Pero os preguntareis, ¿cuándo comienza la historia que nos interesa? La respuesta es sencilla, con los primeros registros de la relación entre medicina y procesos penales. Uno de los textos legales más antiguo y mejor conservado es el Código de Hammurabi que data de 1728 a. C. y que contiene legislación relativa a la práctica de la medicina (3). Volviendo a Egipto, pero en el siglo XVII a. C., se descubrieron textos que describen las diferencias entre distintas heridas por apuñalamiento y profundos conocimientos relacionados con venenos (4). Existen pruebas de que, empleando esta sabiduría, se determinaba la causa de la muerte y se definía si dicha muerte había sido natural o provocada, siendo los médicos llamados a declarar en casos de homicidio o suicidio. A su vez, textos, prácticas e investigaciones relacionadas con el esclarecimiento de la muerte de un individuo han sido registradas en China, Roma y Grecia a lo largo de los siglos (3).

Si bien es cierto que todos estos registros demuestran una relación entre el estudio médico y actos delictivos, en muchos casos no se puede demostrar que esos conocimientos se usasen de forma habitual en procesos legales a fin de resolver crímenes. Esto cambia en Europa en 1553 con la publicación del "Ordenamiento penal del Emperador Carlos V", en el que se estipula que debe emplearse el testimonio de un médico experto para la orientación de los jueces en casos en los que se evalúen lesiones a personas (5). Este conjunto de leyes, considerado el primer cuerpo de derecho penal alemán, fue el germen que propició el inicio y evolución de los sistemas medicolegales en otros países del continente europeo. Aun así, habría que esperar más de 200 años hasta que se definiese por primera vez el término "Medicina Legal" de la mano de François-Emmanuel Fodéré. Este médico francés publicó en 1768 su obra *Traité de Médecine*

Légale (6), en la que emplea dicho término para referirse a la aplicación de conocimientos médicos al derecho. A lo largo del siglo siguiente, diferentes ramas de las ciencias naturales irrumpieron en la aplicación de sus conocimientos a la resolución de procesos criminales. Durante el siglo XIX se lograron grandes avances en los estudios forenses, definiéndose nuevos campos de estudio como la toxicología, la dactiloscopia y la balística forense. Con el surgimiento de estas disciplinas, a finales de siglo, surgió la agrupación de las mismas bajo el término "Ciencias Forenses" (7). Si bien es cierto que todas las disciplinas que conforman las ciencias forenses son de gran interés, en este caso debemos centrar nuestra atención exclusivamente en la genética forense. Por tanto, adentrémonos en su historia y evolución, destacando los distintos avances que han llevado a las pruebas de ADN a convertirse en el análisis de referencia, siempre que la muestra lo permita, en la resolución de casos criminales.

Los seres humanos hemos observado los patrones de herencia durante siglos, pero no fue hasta 1866, con el trabajo de Mendel, que la genética se convirtió en una ciencia formal, con leyes y principios bien establecidos. La genética se podría definir como la ciencia que estudia la herencia biológica, siendo su aplicación al campo forense el uso de esos conocimientos con el objetivo de ayudar en la resolución de casos judiciales. La solicitud de pruebas de ADN es una práctica muy habitual en los tribunales de todo el mundo, consolidándose como una herramienta de gran importancia en el ámbito legal en casos de parentesco, resolución de delitos e identificación humana. Hoy en día, el uso de estas técnicas es casuística habitual en los laboratorios forenses de todo el mundo, sin embargo, su uso lleva acompañándonos menos de 40 años. Lo lógico es preguntarnos, ¿dónde empezó todo? Y, como en muchas otras ocasiones, la respuesta a esta pregunta se fraguó a partir de otra disciplina, ocurriendo años antes de que se propusiese la palabra "genética" (1906) (8).

Si queremos hablar sobre los orígenes de la genética forense, debemos retroceder casi 100 años antes de su nacimiento. Esta disciplina surge a partir de algo con lo que está intrínsecamente relacionado, la evolución. La evolución de una disciplina en otra, gracias a los avances y nuevos descubrimientos en el campo de la biología, siendo el punto de partida la hemogenética forense. Por tanto, por sorprendente que parezca, la historia de la genética forense comienza 53 años antes del descubrimiento de la estructura de la molécula de ADN, remontándonos a 1900 y al descubrimiento del primer polimorfismo genético a manos de Karl Landsteiner (9,10), el grupo sanguíneo AB0. Diez años después de dicho descubrimiento, se demostró su herencia mendeliana (10) y, a partir de ese momento, se caracterizaron más antígenos eritrocitarios, proteínas séricas y enzimas eritrocitarias con el objetivo de analizarlos de forma combinada creando perfiles con mayor poder de discriminación (11). Estamos ante el nacimiento de una nueva disciplina, la hemogenética forense, que sólo tardó 15 años, tras su primera aplicación en la resolución de un caso de parentesco (1915), en convertirse en una técnica estándar en los laboratorios forenses (11). Aunque la combinación de los marcadores enumerados anteriormente proporcionaba una individualización considerablemente alta, las limitaciones de estos análisis lastraban en gran medida su uso en muchos de los escenarios a los que nos enfrentamos en este campo. La baja cantidad de muestra, los efectos de degradación y la existencia de restos biológicos como pelos y restos óseos, limitaba en gran medida la

aplicabilidad de estas técnicas en los escenarios forenses más frecuentes. Por tanto, su uso estaba restringido a condiciones ideales que pocas veces podemos disfrutar en el día a día del laboratorio. En ese momento no se sabía, pero nos encontrábamos a las puertas de uno de los mayores descubrimientos científicos del siglo XX, el cual crearía una disciplina autónoma que emplearía las bases de la hemogenética y revolucionaría las Ciencias Forenses.

1.1.1 Descubrimiento del ADN

Tras el redescubrimiento de las leyes de Mendel en 1900 y el descubrimiento de los ácidos nucleicos en 1910, previamente aislados y llamados nucleínas en 1869, se comenzó a gestar la mayor revolución molecular del siglo XX (12). La cascada de hallazgos producidos a lo largo de los años posteriores sentó las bases de la genética: distribución cromosómica, herencia, ley de Hardy-Weinberg, creación del primer mapa genético, recombinación genética y codificación de proteínas. Todos estos descubrimientos son la antesala de la era del ADN que comienza en 1944 cuando se aísla dicha molécula (13). Será solo nueve años después cuando se produzca el mayor avance en biología molecular del siglo, el descubrimiento de la estructura del ADN. Gracias a la obtención de la primera imagen del ADN por difracción de rayos X obtenida por Rosalind Franklin (14), Watson y Crick definen la doble hélice de ADN en 1953 (15) consiguiendo el Premio Nobel de Medicina en 1962. Con su estructura definida, los estudios de la molécula de ADN avanzan a pasos agigantados, proporcionando un mayor conocimiento en relación con su funcionamiento molecular y su herencia. Gracias a esos progresos se detectó la presencia de secuencias específicas de ADN (16) y se determinó por primera vez la secuencia nucleotídica de dicha molécula (17). Estos estudios sirven como impulso para el desarrollo de nuevas técnicas de análisis y, a su vez, se definen nuevas aplicaciones para la genética molecular, siendo una de ellas su aplicación al campo forense.

Es durante esta vorágine de desarrollo científico cuando se empieza a gestar el nacimiento de la genética forense. Una de las bases de esta disciplina se define en los años 40 por Ford (18) estableciendo el término "*polimorfismo*", siendo descubierto en la molécula de ADN, 40 años después, el primer locus altamente polimórfico, XCH4A-rHsl8 (19). Con todas las piezas sobre el tablero, la genética forense sólo necesitó cinco años para definirse como una disciplina autónoma y emplear los conocimientos descubiertos hasta la fecha para irrumpir en los juzgados de todo el mundo. Por tanto, en 1985, de la mano de Alec Jeffreys y sus colaboradores, se mostraría la gran relevancia que tendría el uso de polimorfismos de ADN al campo forense (20). En 1984 se identificaron regiones repetitivas no codificantes de 33 pares de bases (bp) en el gen de la mioglobina (21). Gracias a esta primera observación, se identifican, distribuidas a lo largo del genoma, regiones repetitivas formadas por repeticiones en tándem de 10-15 bp que se observaban un número de veces variable entre individuos. Esta variabilidad suscita gran interés entre los investigadores, destacando el gran potencial que tiene como herramienta de identificación en el ámbito forense, definiéndose así el término "*DNA fingerprint*" (o "*huella genética*" en español). Estas regiones recibieron el nombre de minisatélites, describiéndolos como regiones hipervariables y demostrándose que la probabilidad de que dos personas tengan el mismo número de repeticiones en un conjunto de minisatélites disminuye a medida que

aumenta el número de marcadores analizados (22). El impacto de este descubrimiento no tardaría en permear hacia el ámbito judicial, siendo empleado exitosamente en la resolución de un caso de filiación en Reino Unido ese mismo año (1985) y, un año después, se aplicaría por primera vez en un caso criminal, exonerando a un individuo e identificando al perpetrador de dos violaciones y asesinatos en Leicestershire, Inglaterra. A partir de este momento, las pruebas de ADN comenzaron a ser empleadas en tribunales de todo el mundo, llegando a España en 1989 (Sumario 1/89, Audiencia de A Coruña).

Pero esta nueva técnica, aunque muy útil y esperanzadora, mostró rápido sus costuras, siendo en ocasiones sus limitaciones inhabilitantes en causas judiciales. En Estados Unidos en 1989 (People v. Castro, 144 Misc, 2d 956) no se admitieron las pruebas de ADN en un proceso judicial de homicidio al no poder obtenerse una reproducción exacta de su resultado (regla Frye). Pero ¿por qué esta técnica denominada RFLP (*Restriction Fragment Length Polymorphism*) era tan difícil de estandarizar? La respuesta se encuentra en el tipo de sondas empleadas en este análisis conocidas como sondas multi locus o MLPs que, como su nombre indica, hibridan con múltiples *loci* repartidos por el genoma. Por tanto, tras el revelado, empleando *Southern blot*, se obtenía un conjunto de entre 10 y 20 bandas por individuo generando así, la llamada "*huella genética*". Esta metodología hacía que la estandarización y comparación de datos empleando esta técnica fuese muy compleja y difícil de llevar a cabo. Con el objetivo de solucionar este inconveniente, se diseñaron sondas unilocus o SLP (*Single Locus Probes*) que permitían detectar una región específica del genoma facilitando la interpretación y comparación de resultados entre laboratorios y entre réplicas (23). Aun solventando el problema de la estandarización, ambas técnicas estaban muy limitadas a la hora de enfrentar a los demonios que siempre acompañan los análisis forenses; baja cantidad y calidad de ADN resultado de procesos de degradación. Estos problemas limitaban en gran medida la aplicación de los análisis de ADN en muchos de los casos criminales, pero no tardaría mucho en desarrollarse una técnica que impulsaría el crecimiento de la genética forense hasta convertirse en uno de los pilares de las ciencias forenses.

Sería en 1986 cuando se definió la técnica que revolucionaría los análisis de ADN y permitiría evaluar y estudiar esta molécula con mayor facilidad. De la mano de Kary Mullis se detallaría la reacción en cadena de la polimerasa (PCR, *Polymerase Chain Reaction*) que permitía amplificar de manera exponencial regiones específicas de ADN (24). La PCR revolucionó la biología molecular al permitir generar millones de copias de una secuencia especifica de ADN a partir de una pequeña cantidad de dicha molécula. Teniendo en cuenta lo frecuente que era encontrarse con una limitada cantidad y calidad de muestra en casos forenses, esta metodología suscitó mucho interés en el campo, dejando obsoleta rápidamente a la RFLP. La introducción de la PCR en el ámbito forense no solo solventó los problemas de sensibilidad de las técnicas anteriores, sino que permitió por primera vez, analizar restos biológicos que antes eran inviables, por lo que empezaron a analizarse pequeñas manchas de vestigios biológicos, pelos, colillas y hasta restos cadavéricos altamente degradados (25). El análisis de estas nuevas muestras abrió las puertas a investigaciones no contempladas anteriormente como, por ejemplo, paternidades cadavéricas, identificación de personas desaparecidas y estudios

prenatales. Pero el avance tecnológico por sí solo no iba a ser suficiente para que la genética forense alcanzase el siguiente nivel. Por muy potente que fuese la técnica, se necesitaban marcadores que fuesen acorde con las necesidades y limitaciones del ámbito. En aquel momento, el análisis de los minisatélites limitaba su uso en muestras críticas, siendo la mayoría de los resultados no concluyentes, debido al gran tamaño de los marcadores analizados (5 a 10 kb). Pero esto iba a cambiar completamente con el descubrimiento de un nuevo tipo de marcadores, los denominados microsatélites o STRs (*Short Tandem Repeats*) cuyas repeticiones en tándem presentaban tamaños de entre 2 a 6 bp. Estos polimorfismos fueron identificados en 1989 (26) y adoptados rápidamente en casuística forense, siendo empleados para la identificación de restos cadavéricos en casos judiciales en 1991 (27) y 1992 (28). Como el interés en estos nuevos marcadores creció sustancialmente, también lo hizo el número de STRs caracterizados a lo largo de los años, tanto en cromosomas autosómicos como sexuales (29–31). Por tanto, fue necesario establecer criterios a cumplir por los marcadores de interés forense, seleccionando sólo aquellos que presentasen alta heterocigosidad, bajo nivel de *stutter*, robustez y baja tasa de mutación (32). Gracias a toda esta atención sobre el tema, no se tardó mucho en diseñar reacciones de PCR que permitían analizar simultáneamente varios STRs, conocidas como reacciones de PCR en *multiplex* (33–35). Con este tipo de análisis se conseguía una reducción en el tiempo de procesamiento y cantidad de muestra necesaria, generando a mayor número de *loci* analizados, un mayor poder de discriminación. El uso de estos marcadores siguió ampliándose, convirtiéndose en poco tiempo en los marcadores de referencia en genética forense, marcadores que hoy en día son el principal medio utilizado en la resolución de casos criminales, de parentesco y de identificación. Motivada por las capacidades y ventajas que suponían los STRs, la comunidad científica no tardaría en generar bases de datos de ADN, basadas en STRs, convirtiéndose en herramientas valiosas para la justicia y las investigaciones forenses. La primera base de datos creada con fines forenses, llamada *UK National DNA Database* (NDNAD) fue establecida en Inglaterra en 1995, estando compuesta por perfiles genéticos con seis marcadores STR. Ésta y otras bases de datos que se generaron posteriormente fueron ampliando el número de marcadores, incluyendo aquellos aceptados internacionalmente por la comunidad científica y que, por tanto, se iban estableciendo como los marcadores de referencia en casuística forense. Si bien es cierto que esta herramienta puede ser de gran ayuda en la resolución de casos criminales, también ha suscitado grandes dilemas éticos en relación con los límites de su uso (36). Estas bases de datos están compuestas por información individualizante que proviene de personas que, a través de un consentimiento, ha decidido cederla con el fin de resolver preguntas muy concretas. El problema se genera cuando esa información se usa con fines no contemplados en ese acuerdo. Y es que, como el tío Ben dijo una vez, parafraseando el antiguo adagio del siglo I a. C. asociado a la espada de Damocles: *"un gran poder conlleva una gran responsabilidad"*.

Con una metodología de amplificación y unos marcadores en auge, el siguiente avance se produjo en el método de análisis de los fragmentos amplificados. En poco tiempo, los geles manuales quedaron obsoletos dando paso a secuenciadores automáticos que pulieron el método de Sanger (17) y lo convirtieron en el principal sistema de detección. La incorporación de

sistemas de marcaje por fluorescencia propició el aumento del número de marcadores por *multiplex*, ya que se empleaban fluorocromos diferentes para marcar los cebadores de *loci* que presentaban el mismo tamaño de fragmento. En los últimos años, las nuevas plataformas de secuenciación masiva en paralelo (MPS, *Massive Parallel Sequencing*) han comenzado a tener un mayor protagonismo en los laboratorios forenses, sobre todo en la parte de investigación. Estas plataformas se basan en la secuenciación simultánea de los fragmentos amplificados tras la PCR, detectando las señales de cada fragmento de ADN secuenciado de forma individual. Estas nuevas tecnologías, un poco difíciles de adaptar a casuística forense por el momento, presentan grandes ventajas con respecto a su hermana mayor, la electroforesis capilar. Entre dichas ventajas destaca la capacidad de analizar un número de marcadores mucho más elevado, existiendo en la actualidad paneles como el *Precision ID GlobalFiler NGS STR* (37) o el *FORCE* panel (38) compuesto, este último, por 5422 marcadores, proporcionando la posibilidad de combinar marcadores con distintas aplicaciones.

A lo largo de todos estos años de avances, se han identificado diferentes tipos de polimorfismos de gran relevancia para la genética forense tanto en cromosomas autosómicos, sexuales y ADN mitocondrial. Estos descubrimientos surgen del trabajo conjunto de la comunidad científica y de proyectos como el Genoma Humano (39) que han permitido ampliar nuestro conocimiento en relación con esta pequeña molécula que porta nuestra información genética y que alberga tantos misterios. Estos nuevos polimorfismos vinieron acompañados de nuevas aplicaciones y herramientas de gran utilidad para la genética forense, ampliando las capacidades del campo y proporcionándonos una era dorada para las ciencias forenses.

1.2. Polimorfismos del ADN

El ser humano se ha definido gracias a la variabilidad genética, dando lugar esta mutabilidad identificadora al surgimiento de la rama forense. Esta disciplina está intrínsecamente relacionada con esa individualidad que permite identificar al donante de un vestigio biológico, empleando posiciones de nuestro ADN que presentan múltiples alelos en la población. ¿Y dónde se encuentra esta variabilidad en nuestro genoma? Para contestar a esta pregunta debemos conocer un poco nuestro ADN. Las secuencias que componen el genoma humano, de forma muy general, pueden dividirse en dos tipos, definidos por su capacidad para transcribirse a proteínas. Por tanto, se diferencia entre ADN expresivo o codificante y ADN no codificante, siendo este último el que no es transcrito a proteínas. Un error habitual es considerar que la falta de esta capacidad está asociada a no funcionalidad, pero muchas regiones de ADN no codificante tienen funciones de regulación génica. Estas secuencias presentan una característica importante y es que una alta proporción de este material genético es ADN repetitivo que, al no estar sujeto a presión selectiva intensa, admite mayores niveles de variación en comparación con el ADN codificante. Por este motivo, en estas secuencias encontraremos los marcadores, definidos como polimorfismos en 1940, que conforman la piedra angular de la genética forense. Pero ¿cuándo se clasifica como polimorfismo una posición o secuencia variable? Teniendo en cuenta la recomendación de analizar 100 o más individuos, se considera polimórfica cualquier variante cuyo alelo más común tenga una frecuencia en la población estudiada inferior al 99%

(40). Tras definirse criterios de selección para marcadores aplicables a la genética forense, se han seleccionado aquellos polimorfismos, de forma individual o combinada, que proporcionasen mayor informatividad en los diferentes usos aplicables en esta amplia disciplina.

Entre los polimorfismos de mayor uso en genética forense se encuentran los polimorfismos de longitud, que consisten en repeticiones de fragmentos de ADN de número variable (VNTR, *Variable Number of Tandem Repeats*), y polimorfismos de secuencia, que hacen referencia a aquellas variaciones que ocurren en un solo nucleótido (SNP, *Single Nucleotide Polymorphism*). Estos polimorfismos han sido identificados y empleados en la resolución de múltiples casos criminales a lo largo del tiempo. Gracias al descubrimiento de nuevos marcadores no sólo se ha conseguido mejorar el poder de discriminación de los análisis, sino que también, se han desarrollado nuevas herramientas que proporcionan información de gran relevancia en investigaciones judiciales. Por tanto, es necesario conocer los distintos tipos de polimorfismos empleados en genética forense, sus aplicaciones y el papel de los cromosomas sexuales y del ADN mitocondrial.

1.2.1 *Short Tandem Repeats*

Los *Short Tandem Repeats* (STRs), también conocidos como microsatélites, son los principales marcadores de rutina en genética forense. Se definen como repeticiones cortas de ADN en tándem formadas por unidades repetitivas de entre 2 a 6 bp, formando conjuntos de hasta 100 bp (41). Estos marcadores se definen como polimorfismos de longitud, empleándose el número de repeticiones del motivo repetitivo para la identificación los alelos existentes, siendo habitualmente de entre 10 y 30 repeticiones (42), un ejemplo de este tipo de marcadores se esquematiza en la Figura 1.

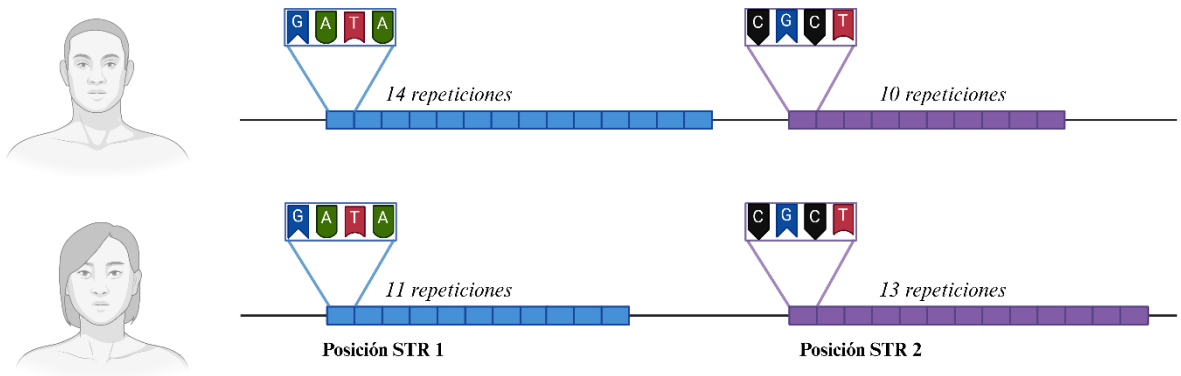


Figura 1. Representación de dos marcadores STR compuestos por un motivo de repetición diferente y que presentan un número de repeticiones distintas entre dos individuos. Figura creada con BioRender.

Gracias a su pequeño tamaño su descubrimiento supuso una gran ventaja, opacando rápidamente a sus antecesores los minisatélites, cuyo tamaño total variaba entre 300 bp y 20 Kb. Esta diferencia sustancial de tamaño supondría un gran avance en los análisis forenses permitiendo, por primera vez, obtener resultados de muestras anteriormente inanalizables. A su vez, estos marcadores difieren también en su distribución en el genoma humano, siendo mucho

más abundantes los microsatélites que los minisatélites. Los STRs representan el 3% del genoma humano existiendo aproximadamente 700.000 *loci* (43), observándose con una frecuencia de 1 cada 2.000 bp, presentando una distribución y densidad que varía entre cromosomas (41). A su vez, es importante destacar que solo el 8% de los STRs se encuentran en regiones codificantes, presentándose predominantemente en regiones no codificantes (44). Gracias a que la mayoría están localizados en estas regiones del ADN, los avances tecnológicos no se ven limitados en los países en los que su legislación no permite el uso de ADN codificante para la identificación de personas, como por ejemplo España (45).

Como se puede observar, las unidades de repetición de los STRs son las principales protagonistas al hablar de estos marcadores, ya que los definen, nombran y clasifican. La clasificación de los STRs se realiza teniendo en cuenta la longitud y la estructura del motivo de repetición. La longitud hace referencia al número de nucleótidos que posee el motivo de repetición, por tanto, podemos clasificarlos en mono-, di-, tri-, tetra-, penta- y hexanucleotídicos. Es importante destacar que, en el genoma humano, los STRs más comunes son los compuestos por unidades dinucleotídicas, esto se debe a que el tamaño del motivo está relacionado con la frecuencia del STR. A medida que aumenta el tamaño de la unidad de repetición disminuye la ocurrencia en el genoma (46,47). Entre ellos, los más usados en identificación humana son los STRs tetranucleotídicos, ya que generan un menor número de bandas *stutter* en comparación con los di- y trinucleotídicos. Una menor probabilidad de artefactos genera una ventaja que facilita la interpretación de dichos marcadores, principalmente, en mezclas de ADN, muestras muy habituales en genética forense. A parte de la longitud de la unidad de repetición también se evalúa su estructura para realizar una correcta clasificación. Los motivos de repetición de un STR pueden ser idénticos entre sí, con unidades incompletas o presentando unidades de repetición distintas, teniendo esto en cuenta se pueden clasificar como (48):

- STRs simples: unidades de repetición idénticas en longitud y secuencia.
- STRs simples con alelos no consenso: alguna unidad de repetición incompleta.
- STRs compuestos: dos o más tipos de unidades de repetición con secuencias diferentes.
- STRs compuestos con alelos no consenso: STRs compuestos con alguna unidad de repetición incompleta.
- STRs complejos: varios bloques de repetición con longitud variable.

En este caso, a diferencia de la longitud del STR, la complejidad estructural no limita la selección de estos marcadores. La selección de STRs para uso forense depende principalmente del poder de identificación de los *loci*, que no estén ligados, que presenten una baja tasa de mutación, bajo número de artefactos y que sean validados para su uso en forense.

Los STRs se han convertido en los marcadores de referencia en genética forense, la automatización de su genotipado y el fácil análisis e interpretación de los resultados los han situado como punta de lanza para todos los laboratorios forenses del mundo. Pero un fácil

análisis no justifica su popularidad. Estos marcadores presentan importantes ventajas que les confieren merecidamente el estatus que presentan. Los STRs destacan por ser altamente polimórficos, con alto poder de discriminación debido a su naturaleza multi alélica, presentan tamaños de amplicón que varían entre 100 a 400 bp y pueden ser amplificados en *multiplex*. Estas características no solo los sitúan como marcadores de referencia para la identificación humana, sino que también les confieren cierta utilidad en la inferencia de la ancestralidad de un individuo gracias a que la distribución alélica de los STRs es variable en las distintas poblaciones (49).

Gracias a todas estas ventajas, los STRs llevan usándose en casos forenses desde 1990 (27,28) y no parece que vayan a irse de nuestras vidas en mucho tiempo. Si bien es cierto que su uso va más allá del campo forense, siendo empleados en diagnóstico clínico y mapeo genético, la genética forense ha crecido y evolucionado gracias y alrededor de ellos. Con el paso del tiempo, se han desarrollado paneles de STRs con un alto grado de discriminación e implementado nuevas técnicas que permiten un análisis rápido, fácil de interpretar y partiendo de ADN de poca cantidad y calidad. La tecnología avanza y los análisis de estos marcadores deben evolucionar con ella, por lo que se están desarrollado nuevos paneles de STRs analizados con tecnologías de secuenciación masiva en paralelo que puedan ofrecer un mayor poder de discriminación. Aún con todo esto, existen otros marcadores en genética forense que complementan a los STRs o que aportan otro tipo de información que puede ser de mucha utilidad para la resolución de investigaciones criminales.

1.2.2 *Single Nucleotide Polymorphisms*

Los *Single Nucleotide Polymorphisms* (SNPs) son variaciones en el ADN que afectan a una única base de la secuencia del genoma, como se muestra en la Figura 2. Pero no todas las mutaciones pueden considerarse SNPs. Para que una variación pueda entrar en este selecto grupo, que no se puede considerar pequeño, debe darse en al menos el 1% de la población, siendo consideradas mutaciones puntuales aquellas que no cumplen este requisito (50). Lo cierto es que estos marcadores no pueden considerarse escasos, ya que están ampliamente distribuidos en el genoma humano y representan casi el 90% de la variación genética humana (11), presentándose en 1 de cada 500 bp (51). Los SNPs son mutaciones que han sido exitosas evolutivamente fijándose en una parte importante de la población humana, aportando información individualizante, poblacional, fenotípica y relacionada con enfermedades.

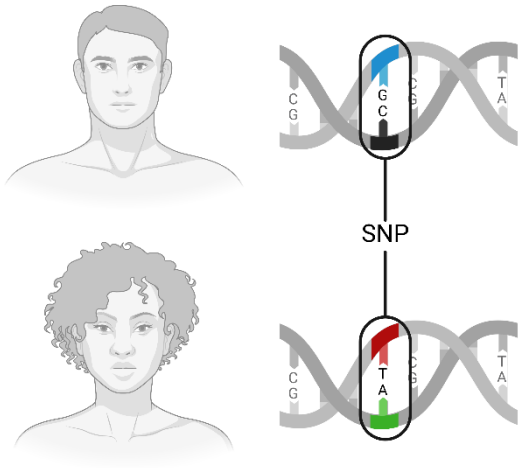


Figura 2. Representación de un marcador SNP. Figura creada con BioRender.

Un capítulo importante en la historia de los SNPs comienza con el inicio, en 1984, del Proyecto Genoma Humano (HGP) (52). Gracias a este trabajo, y proyectos derivados de éste, se reveló la existencia de estos polimorfismos que se encontraban con una densidad muy alta en el genoma. A partir de este momento, estos marcadores despertaron el interés de la genética clínica y forense, desarrollándose tecnologías que permitían analizar de forma automática un gran número de SNPs, como por ejemplo un *microchip* analizado por Wang et. al (53) con el que estudiaron miles de SNPs simultáneamente. Todos estos análisis fueron configurando las ventajas y desventajas que estos marcadores presentaban en comparación con los STRs. En primer lugar, para realizar la comparativa correctamente, debemos destacar una característica importante de los SNPs, son marcadores bialélicos. Esta particularidad, en contraste con los STRs que son marcadores multialélicos, conlleva que los SNPs sean menos informativos (54), requiriendo el estudio de un mayor número de marcadores para alcanzar niveles de discriminación similares a los que proporcionan los STRs (55). Para intentar compensar esta informatividad reducida se recomienda emplear 4,2 SNPs (con frecuencias alélicas que oscilen alrededor de 0,5) por cada STR (56).

Si bien es cierto que hay que tener en cuenta esta desventaja, las ventajas de los SNPs compensan esta limitación con creces. En primer lugar, los productos obtenidos tras el análisis de un SNP pueden tener un tamaño inferior a 100 bp, pudiendo hacer frente a muestras degradadas con mayor facilidad que los STRs (57,58). Por otro lado, los SNPs pueden agruparse en una misma reacción en mayor grado, siendo procesados y analizados de forma más automática que los polimorfismos de longitud. Otra ventaja de estos marcadores es que no presentan bandas *stutter* haciendo más fácil la interpretación de los alelos analizados. A su vez, estos marcadores presentan una ventaja por sí mismos, pueden emplearse como marcadores complementarios en casos en los que los marcadores de rutina no sean lo suficientemente informativos. Por ejemplo, cuando se trabaja con relaciones de parentesco lejanas, el análisis de marcadores adicionales es necesario para alcanzar una adecuada razón de verosimilitud. Los SNPs, presentando una menor tasa de mutación (10^{-8}) en comparación con los STRs (10^{-3}), son marcadores más estables a lo largo de las generaciones, proporcionando información adicional a los STRs cuando en éstos se detecta la presencia de eventos mutacionales (59).

A parte de este uso complementario, el análisis de SNPs presenta una gran ventaja en investigaciones criminales, proporcionando información adicional sobre el donante de un vestigio biológico. Cuando nos enfrentamos a situaciones en las que no hay sospechoso o el perfil genético obtenido no presenta coincidencias con las bases de datos de ADN, la inferencia de la ancestralidad o la predicción de características físicas, pueden ser de gran ayuda para identificar al donante de una muestra biológica. La predicción del origen biogeográfico se lleva a cabo empleando SNPs clasificados como *Ancestry-Informative-Markers* (AIMs). En esta aplicación, la ventaja de los SNPs sobre los STRs está asociada a la tasa de mutación de estos marcadores (10^{-8} vs 10^{-3}, respectivamente). Como comentamos previamente existe una diferencia sustancial provocando que los SNPs estén con mayor probabilidad fijados a una población (60). Actualmente este tipo de análisis están validados a nivel forense, presentando una alta eficacia y ofreciendo elevadas probabilidades de predicción. La importancia de estos marcadores y su uso en el campo forense quedó patentada en muchas ocasiones, siendo una de las más reconocidas, su uso para identificar el origen biogeográfico de perfiles no identificados relacionados con el atentado del 11-M en Madrid (61). Por otro lado, gracias a la generación de un conocimiento más profundo del genoma, se identificaron posiciones correlacionadas con características fenotípicas. La predicción de características visibles externas del donante de la muestra ha suscitado gran interés forense, generándose paneles de marcadores que permiten predecir color de pelo, color de ojos y color de piel (62), entre otros. Actualmente la mirada está centrada en marcadores fenotípicos que permitan predecir la morfología facial, con el objetivo de proporcionar un retrato molecular a partir del ADN de una muestra biológica (63). Estos estudios se encuentran limitados, por el momento, por la difícil búsqueda de marcadores correlacionados con características faciales y la complejidad de los modelos de predicción matemáticos con los que alcanzar una reconstrucción facial fiable.

El trabajo entre laboratorios de todo el mundo, al participar en proyectos internacionales, ha propiciado la creación y validación de paneles de predicción de ancestralidad y características fenotípicas de gran interés forense (64–66), siendo su aplicación aceptada a nivel

judicial y solicitada actualmente por los investigadores en los casos criminales que lo requieren. Todo este descubrimiento de marcadores, validación de paneles y nuevas tecnologías, así como generación de modelos de predicción y nuevas aplicaciones, son fruto de un esfuerzo conjunto de la comunidad científica forense. La mejor forma de avanzar es unir recursos y conocimientos, como dijo Gayo Salustio Crispo "*concordia res parvae crescunt*" (las pequeñas cosas florecen de la concordia), frase adaptada por la República Holandesa al famoso dicho *"La unión hace la fuerza"*.

1.2.3 Microhaplotipos

Los proyectos que investigan el genoma humano han permitido llevar a cabo un mayor número de estudios con el objetivo de desentrañar la evolución humana y la estructura poblacional. Gracias a todos estos trabajos se identificaron estructuras en bloque de *loci* asociados (67). Si bien es cierto que su estudio ha despuntado en los últimos años, dichas estructuras fueron definidas en los años 60 como "haplotipos" (68). Los trabajos de proyecto genoma humano (47) y 1000 *Genome Project* (69) propiciaron el estudio de este nuevo tipo de *loci* de interés forense, identificándose los microhaplotipos (MHs) (70,71). Estos marcadores moleculares, como se puede observar en la Figura 3, comprenden pequeños segmentos de ADN (<300 bp) que presentan dos o más SNPs estrechamente ligados con tres o más combinaciones alélicas. Los microhaplotipos han demostrado ser marcadores prometedores y versátiles que pueden emplearse con distintos fines en la resolución de casos criminales. En los últimos años se ha probado su aplicabilidad con el fin de extraer información relacionada con inferencia de la ancestralidad (72), deconvolución de mezclas (73), identificación humana (74) y pruebas de parentesco (73).

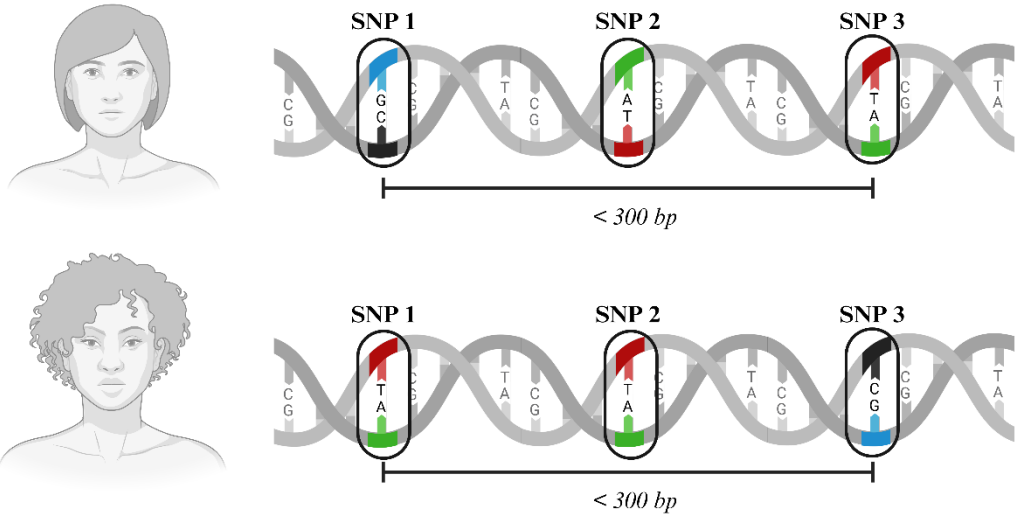


Figura 3. Representación de un conjunto de tres SNPs que componen un microhaplotipo. Figura creada con BioRender.

Debido a su estructura, los microhaplotipos presentan las mismas ventajas que los SNPs mitigando alguna de las desventajas de estos marcadores. Al estar conformados por un conjunto de SNPs, proporcionan más información y poder de discriminación que un SNP por sí solo.

Aun así, al compararlos en este aspecto con los STRs, los microhaplotipos presentan generalmente un menor poder de discriminación. Esto se debe a que los MHs presentan un menor número de alelos, siendo necesario un mayor número de marcadores para superar la informatividad de un STR. Por otro lado, la ausencia de secuencias repetidas en tándem en estos nuevos marcadores elude la ocurrencia de *slippage*, error que provoca la generación de artefactos y que suele ser común en los análisis con STRs. Por tanto, los microhaplotipos no generan bandas *stutter* facilitando el análisis de mezclas de ADN. A su vez, los MHs presentan una menor tasa de mutación, confiriéndole una ventaja clara sobre los STRs en análisis de parentesco (73). Si bien es cierto que estos nuevos marcadores moleculares presentan un gran potencial, para algunos laboratorios su uso puede verse limitado por la necesidad de realizarlos empleando tecnologías de MPS (71). De todas formas, esta limitación podría verse solventada con el uso de MinION, la tecnología de *Oxford Nanopore Technologies* de secuenciación portátil en tiempo real (75).

1.2.4 Polimorfismo del cromosoma X

En ciertos escenarios forenses es de gran utilidad identificar el linaje familiar tanto por vía paterna como materna, analizado los polimorfismos descritos en los apartados anteriores en el cromosoma X, cromosoma Y y en el ADN mitocondrial. Esta información genética es heredada de uno de los dos progenitores, herencia monoparental, sin la recombinación que ocurre en el ADN autosómico, confiriéndoles ciertas ventajas y desventajas con respecto a los marcadores analizados en el ADN de herencia biparental. En relación con los cromosomas sexuales, el par cromosómico 23, determina el sexo de los humanos con la combinación del cromosoma X y el cromosoma Y, siendo homogamético el genotipo femenino (XX) y heterogamético el genotipo masculino (XY). Al hablar de herencia es importante destacar el modelo híbrido que presenta el cromosoma X, transmitiendo el padre a su descendencia femenina su cromosoma X prácticamente íntegro y la madre una recombinación de sus dos cromosomas X. Esta dualidad sitúa a este cromosoma en un punto intermedio entre los cromosomas autosómicos y los genomas de herencia uniparental (cromosoma Y y ADN mitocondrial) proporcionándole características exclusivas (76).

El uso de polimorfismos del cromosoma X ha demostrado tener gran potencial tanto como marcadores de apoyo o individualmente en casos de identificación humana, parentesco y genética de poblaciones. La identificación de STRs en este cromosoma se remonta a principios de los 90 (29,77), desarrollándose con el paso del tiempo *multiplexes* altamente estandarizados. Las investigaciones relacionadas con este cromosoma se han visto reducidas con el paso del tiempo, pero en los últimos años se siguen describiendo nuevos marcadores de interés forense (78). Actualmente, los estudios relacionados con estos marcadores son principalmente poblacionales (79–81), pero la falta de bases de datos para la accesibilidad y uso de frecuencias alélicas de los X-STRs o haplotipos limita su utilización en gran medida (82).

El uso forense de marcadores de cromosomas sexuales en los casos de parentesco se ve principalmente limitado, irónicamente, por su característica principal, su modelo de herencia.

En el caso de los X-STRs, su uso se reduce principalmente al estudio de relaciones de parentesco padre – hija (en ausencia de madre), abuela – nieta, medias hermanas con padre común y paternidades en casos de incesto (83). A su vez, el análisis del cromosoma X puede ser de gran utilidad a la hora de excluir a parientes cercanos del verdadero padre (84). Estos marcadores se ven actualmente relegados a un segundo plano, no siendo utilizados de forma muy común en rutina forense. Esto podría deberse al decaimiento de los últimos años del interés forense en estos marcadores, derivado de los usos tan específicos que presentan y la difícil interpretación de la naturaleza híbrida de su modelo de herencia (85). Este decrecimiento del interés con respecto al cromosoma X no va de la mano con su homólogo masculino, siendo el cromosoma Y, de gran interés e importancia en casuística forense.

1.2.5 Polimorfismos del cromosoma Y

El uso de polimorfismos del cromosoma Y se ha convertido en algo rutinario en cualquier laboratorio de genética forense. Su uso tan obvio y necesario en el ámbito forense, tardó en identificarse tras el descubrimiento del primer polimorfismo genético en el cromosoma Y humano en 1985 (86). Casi 10 años después de este suceso se descubrió el primer Y-STR (87) y se empleó su análisis por primera vez, en 1992, en un caso judicial para identificar a un asesino varón en una mezcla de ADN (88). Esto puso el foco de interés sobre estos marcadores y su, en aquella época, posible potencial para la resolución de mezclas. En muchos casos, el componente minoritario de una mezcla era difícil o imposible de detectar, por tanto, en casos de violación, en los que mayoritariamente el componente minoritario es masculino, el uso de Y-STRs comenzó a ganar importancia (88). Esto propició el desarrollo de diversos estudios que evaluaron distintos grados de mezclas de ADN masculino y femenino (89). Con estos estudios se apuntaló el uso de estos marcadores en casos de criminalística despuntando con la comercialización, tardía (2003 y 2005), de los primeros *kits* comerciales de *multiplex* de Y-STRs (90,91).

No sería muy arriesgado decir que la aplicación anteriormente descrita es la más destacada e importante de estos polimorfismos, pero no podemos olvidarnos de otras que también han sido y son de gran utilidad. Los polimorfismos del cromosoma Y (Y-STRs e Y-SNPs) se han empleado con el objetivo de inferir el origen biogeográfico (92–94), definiéndose haplogrupos específicos del cromosoma Y que mostraban entre ellos una limitada distribución geográfica, lo que los convierte en una herramienta de inferencia de la ancestralidad muy útil. La mejor forma de demostrar la capacidad de esta aplicación es mostrar su uso en un caso judicial. Un ejemplo perfecto podemos encontrarlo explicado con todo lujo de detalles en la revisión publicada por Manfred Kayser en 2017 (95), en la que comenta el caso del asesinato de Marianne Vaatstra. Marianne fue víctima de una violación y asesinato, siendo su cuerpo encontrado con restos biológicos de su asesino sobre él. La muestra de ADN del perpetrador era de buena calidad, pero no se obtenía ninguna coincidencia para los STRs autosómicos analizados. Viendo que todos los sospechosos eran excluidos se decidió cambiar de enfoque, poniéndose en contacto con Peter de Knijff (investigador del Departamento de Genética Humana del Centro Médico de la Universidad de Leiden) y Lutz Roewer (investigadora del

Instituto de Medicina Legal, *Charité University Medicine Berlin*) para analizar Y-STRs e inferir la ancestralidad biogeográfica paterna del asesino. Gracias a estos análisis las investigaciones se centraron en personas del noreste de Europa. Esta información estrechó la búsqueda, pero no se logró identificar al asesino ya que el análisis biogeográfico y el uso del ADN en búsqueda de familiares no estaba permitido por la ley. Pasaron muchos años hasta que, gracias a los Y-STRs y las nuevas legislaciones asociadas a investigaciones forenses, permitieron analizar e identificar la línea paterna del haplotipo determinado y se identificaron dos apellidos relacionados con un ancestro común. Siguiendo esta pista se identificó al perpetrador 14 años después del hecho. Como se puede observar en este caso, los STRs del cromosoma Y pueden ser de gran ayuda en investigaciones policiales no sólo identificando el origen biogeográfico, sino también, en la identificación de la línea paterna del haplogrupo obtenido.

Gracias a casos como éste, el uso de los polimorfismos del cromosoma Y se amplificó, normalizándose tanto técnica como legalmente, dando paso al desarrollo de paneles y bases de datos (YHRD, *Y-Chromosome Haplotype Reference Database* (96)) que facilitasen su uso en casuística forense.

1.2.6 Polimorfismos del ADN mitocondrial

Las mitocondrias son orgánulos celulares que contienen un genoma circular conocido como ADN mitocondrial. Hoy en día su secuencia ha sido secuenciada por completo, observándose ciertas regiones de este genoma que mutan entre 5 y 10 veces más rápido que el ADN nuclear. Estas regiones hipervariables son de gran interés para los estudios de genética forense, advirtiéndose que la mayor cantidad de variaciones de secuencia entre individuos se encuentran en las regiones denominadas: región hipervariable 1 (HV1) y región hipervariable 2 (HV2) (97). Esta variabilidad, su herencia exclusivamente materna y la existencia de miles de mitocondrias por célula confieren a este genoma múltiples aplicaciones en el campo de la genética forense.

El alto número de copias del ADN mitocondrial, su estructura circular y una gran estabilidad, convierten a esta molécula en un objetivo de interés cuando nos enfrentamos a muestras degradadas. Por tanto, se convierte en una herramienta de gran utilidad cuando el ADN nuclear está altamente degradado o es inexistente, como en el análisis de pelos y cabellos sin bulbo. A su vez, su herencia exclusivamente materna lo convierte en un material genético muy útil para estudios de linaje, como se demostró con el análisis de los huesos de la familia Romanov (98). El Zar Nicolas II, fue identificado gracias a la comparación del ADN mitocondrial de los restos óseos encontrados, con el obtenido a partir de dos parientes maternos vivos. El potencial forense de este genoma fue usado por primera vez en un caso de agresión sexual en un tribunal en 1996, tanto en Estados Unidos como en España (caso P.W Ware y Sumario 2/95, Juzgado de 1ª Instancia e Instrucción de Puente Genil, Córdoba, respectivamente). Por último, la alta tasa de mutación del ADN mitocondrial en comparación con el ADN nuclear convierte a este genoma en una herramienta de gran utilidad para el estudio de poblaciones humanas. Por este motivo, se ha empleado el método filogenético para estudiar

y caracterizar la región control del ADN mitocondrial humano, definiéndose haplogrupos de utilidad para la inferencia del origen biogeográfico (99).

Gracias a todos estos polimorfismos identificados en los dos genomas presentes en el cuerpo humano, en pocos años, los análisis de ADN habían cambiado por completo la forma de resolver crímenes. Pero, como dijo Shuden en el libro Elantris de Brandon Sanderson *"En cuanto una revolución consigue su objetivo, otra empieza a planearse"*. Todos estos polimorfismos presentan un talón de Aquiles a la hora de llevar a cabo la identificación del donante de un vestigio biológico. Por muchos marcadores que se genotipen se necesita una muestra de referencia o base de datos con un perfil coincidente para llevar a cabo una identificación. Por tanto, ¿cómo podemos afrontar los casos en los que no hay sospechoso o coincidencia en las bases de datos de ADN disponibles? La respuesta a esta pregunta ya se ha dejado entrever en los apartados anteriores y está relacionada con la revolución propiciada por las herramientas a las que comúnmente nos referimos como "*DNA intelligence tools*". Esta alternativa nace de la identificación de marcadores que permiten aportar información adicional sobre el donante de la muestra (por ejemplo: origen biogeográfico, características físicas y edad). Esa información puede ser de gran utilidad para los investigadores, reduciendo el número de sospechosos u orientando las investigaciones hacia individuos que no se habían tenido en cuenta previamente. A continuación, nos adentraremos en el conjunto de herramientas desarrolladas con este objetivo, descubriendo los últimos avances y nuevos horizontes de la genética forense.

1.3. *DNA intelligence tools*

La recuperación de un perfil genético a partir de un vestigio biológico y su coincidencia con una muestra de referencia convierte a las evidencias de ADN en herramientas extremadamente poderosas para la resolución de crímenes. ¿Pero qué ocurre cuando no existe dicha coincidencia, cuando el perfil de la evidencia obtenido no permite identificar a nadie? En estas situaciones las investigaciones se ven ralentizadas, obligando a los investigadores policiales a seguir otros hilos conectores en busca de nuevas pistas que los guíen hasta el perpetrador. En estas situaciones puede realizarse un cribado en masa en la zona o zonas de interés para el caso, como en el caso anteriormente comentado de Marianne Vaastra o, sin irnos muy lejos, durante el caso de Eva Blanco en España. El éxito de estos cribados no está garantizado y depende en gran medida de que el autor del crimen o sus familiares (como ocurrió en los casos citados) se encuentren en el área de estudio. Por otro lado, no se puede obviar el gran gasto económico que conlleva y los dilemas éticos subyacentes a este tipo de estudios en masa, lo que hace que sean análisis muy difíciles de llevar a cabo (100). Por tanto, podemos encontrarnos con casos que permanecen irresolubles durante muchos años.

Estas limitaciones han estimulado el desarrollo de nuevas herramientas denominadas *DNA intelligence tools* que tienen como principal objetivo lo que se conoce como *Forensic DNA Phenotyping* (FDP) o Fenotipado Forense del ADN. La finalidad de este tipo de estudios se centra en la identificación y aplicación de marcadores que permitan llevar a cabo la predicción

de ciertas características del donante de una muestra biológica. Con ellos se genera información que pretende reducir el número de sospechosos, orientando las investigaciones policiales con el objetivo de obtener una coincidencia entre el perfil recuperado en la escena de un crimen y el perpetrador (101). La resolución de crímenes no es la única aplicación forense de las *DNA intelligence tools*, estos análisis pueden ser de gran utilidad en la identificación de personas desaparecidas, cuando no existen muestras *ante mortem* de la persona o familiares cercanos con los que comparar. Podemos pensar en los modelos de predicción que se generan con estos marcadores, como testigos biológicos, siendo éstos mucho más precisos que los testigos oculares, los cuales se ha demostrado que albergan poca fiabilidad (102). Antes de la era genómica y la estandarización de las pruebas de ADN en el ámbito judicial, la única fuente de este tipo de información eran los testigos oculares. Pero ¿cómo de precisa puede ser esa descripción? ¿cuántas personas son capaces de identificar al perpetrador de un crimen sin estar sugestionadas o limitadas si el delito se cometió en situaciones de baja visibilidad? Todas estas dudas razonables nos hacen cuestionarnos si este tipo de descripciones o identificaciones son más contraproducentes de lo que imaginamos. Un dato muy destacable es el aportado por el *Innocent Project*, donde más del 60% de sus clientes fueron encarcelados erróneamente por culpa de la identificación incorrecta de los testigos (103). A su vez, en el registro nacional de exoneraciones de Estados Unidos se puede comprobar que de los 3.422 casos de exoneración, 935 de ellos presentan identificaciones erróneas de los testigos (104). Con esto en cuenta, la necesidad de la implementación de las técnicas FDP queda patente, prescindiendo de los testigos oculares que pueden proporcionar información incierta o guiar incorrectamente a los investigadores.

Cuando pensamos en características visibles externas (EVC, *Externally Visible Characteristics*), lo primero que se nos viene a la mente es la predicción de morfología facial, color de piel, pelo y ojos, morfología de cabello, calvicie, presencia de pecas, etc. Al hacer esta asociación estamos obviando la presencia entre los FDP de otras aplicaciones de gran interés en el ámbito forense, como es la inferencia del origen biogeográfico o la edad. La predicción de estos rasgos no tiene por qué estar asociada a características externas específicas, pero son consideradas parte de los FDP (101,105,106) ya que proporcionan información que puede ser de gran utilidad en investigaciones policiales. En el caso de la edad, se puede identificar una asociación entre la estimación de la edad cronológica y ciertas características externas, como es el encanecimiento y la calvicie masculina. De todas formas, es necesario tener en cuenta que dichos rasgos pueden producirse en edades tempranas diluyendo la asociación directa, pero proporcionando otro tipo de información de interés. Por ahora dejemos la edad en espera y centrémonos en la predicción del origen biogeográfico y de características externas visibles. La predicción de la edad será tratada de forma extensa en un apartado propio, un lugar que se ha ganado a pulso, al estar esta tesis basada principalmente en ella.

Teniendo en mente el tipo de rasgos que se quieren interrogar, ancestralidad y características físicas, solo falta definir el tipo de marcadores que mejor se ajuste a dicho análisis, siendo en este caso los SNPs. Gracias a las características de estos polimorfismos, comentadas anteriormente, se llevaron a cabo estudios de asociación de genoma completo con

el objetivo de identificar marcadores estadísticamente asociados con estos rasgos (107). A partir de este tipo de estudios, a lo largo de los años, se han identificado conjuntos de SNPs que pueden emplearse en el desarrollo de modelos estadísticos que nos permiten predecir el origen biogeográfico o rasgos físicos de interés forense. Adentrémonos, un poco más, en las características estudiadas y en los paneles desarrollados a lo largo de los últimos años, elementos que hoy en día se usan en casuística forense cuando nos enfrentamos a casos que necesitan dar respuesta a preguntas más específicas.

1.3.1 Inferencia del origen biogeográfico

El origen biogeográfico o ancestralidad se puede definir como la herencia genética que cada individuo porta de sus ancestros. La inferencia del origen biogeográfico (BGA) busca detectar variaciones poblacionales encontradas en un individuo que permitan asociar su origen genético a regiones geográficas concretas. Para poder llevar a cabo este tipo de estudios e inferencias, empleando los conocimientos adquiridos de la variación genómica humana, se seleccionan marcadores que presentan diferencias entre poblaciones. Esta diversidad poblacional del genoma fue estudiada por Rosenberg et. al (108) empleando los datos disponibles a partir del Proyecto de Diversidad del Genoma Humano (HGDP) conocido como HGDP-CEPH. Gracias al estudio de estos datos, Rosenberg identificó agrupaciones definidas de forma continental que consistía en Eurasia, África Subsahariana, Este de Asia, América y Oceanía (K:5), generándose además una división en siete poblaciones (K:7) dividiendo Eurasia en Europa, Oriente Medio y Sur de Asia. Evaluando estas agrupaciones poblacionales se identificó que solo el 4,3%/3,6% (K:5 y K:7, respectivamente) de la variabilidad del genoma humano estaba asociado a diferencias poblacionales. A partir de este punto y empleando AIM-SNPs, por sus ventajas anteriormente definidas, se diseñó en 2003 el primer panel, compuesto por 178 SNPs, específico de uso forense para la predicción de origen biogeográfico (109,110). Gracias a estos estudios se definió una nueva herramienta muy importante en el ámbito forense, desarrollándose, ampliándose y aplicándose con gran rapidez. El uso de estos marcadores ha proporcionado pistas clave en la identificación de asesinos años después de cometerse un crimen. El asesino de Milica van Doorn fue identificado 15 años después al predecirse el origen biogeográfico de los restos de ADN encontrados en la escena del crimen. Esta información permitió reducir el número de sospechosos y el perpetrador fue identificado a través de un cribado en masa (111). Este tipo de información no solo es útil en estos escenarios, a veces puede exculpar a grupos poblacionales que se encuentran en la zona y que durante el transcurso de una investigación están siendo perseguidos injustamente como culpables. Un suceso como este se vivió durante el caso Vaatstra. La población local centró sus acusaciones y amenazas sobre un pequeño grupo de refugiados no europeos que estaban recibiendo asilo en la zona en el momento del asesinato. Al determinar el origen biogeográfico del perpetrador, asociándolo con mayor probabilidad a la población local (de ancestralidad europea), se redujo la persecución pública sobre los refugiados (95). Por último, otro ejemplo del potencial de esta herramienta fueron los análisis llevados a cabo tras el atentado del 11-M en Madrid. En este caso de gran presión y tensión mediática, donde se empezaban a producir acusaciones infundadas, la predicción del origen biogeográfico de restos biológicos asociados a los perpetradores fue clave en su identificación.

Empleando 34 AIM-SNPs autosómicos se estudió la ancestralidad de siete muestras evaluándose una pregunta específica formulada por el juez a cargo de la investigación, definir si los donantes de dichas muestras eran de ancestralidad europea o del norte de África. Tras el análisis y la evaluación estadística de las muestras, se clasificaron tres de ellas como no concluyentes, una de ancestralidad europea y tres de ancestralidad del norte de África (61).

Con el paso de los años se han desarrollado diversos paneles que permiten predecir el origen biogeográfico. Este desarrollo se realizó de la mano de las nuevas tecnologías de MPS que permitieron construir paneles más amplios que permiten diferenciar con mayor precisión un mayor número de poblaciones (66). En la mayoría de los paneles de ancestralidad desarrollados, los SNPs se posicionan como los marcadores de preferencia, pero en los últimos años, los microhaplotipos han demostrado un gran potencial en la inferencia de la ancestralidad y sus ventajas con respecto a los SNPs podría convertirlos, con el tiempo, en los marcadores de preferencia en estos estudios (65). Es importante destacar que comienza a ser una tendencia habitual construir dichos paneles combinando marcadores de origen biogeográfico con otros relacionados con otras aplicaciones, como por ejemplo para predicción de características físicas (106). La predicción de la ancestralidad por sí sola puede ser una herramienta con gran potencial a la hora de resolver un crimen, pero la adición a esa información de características externas visibles como el color de ojos, pelo y piel pueden reducir de forma más significativa el número de sospechosos.

1.3.2 Predicción de características físicas

La predicción de características externas visibles se convirtió en un tema de interés en el ámbito forense a principios de los años 2000 (112), pero su progreso fue, y sigue siendo, lento debido a la complejidad de los rasgos analizados. El análisis y predicción de los EVCs es una rama desafiante debido a que los rasgos a predecir se ven influenciados simultáneamente por multitud de genes y en algunos casos factores ambientales. Estas características son genéticamente muy complejas ya que en muchos casos el gran número de genes que las determinan contribuyen en una pequeña proporción a la definición de dicho rasgo, siendo difícil identificarlos estadísticamente y necesario el uso de paneles con un elevado número de marcadores (107). Si bien es cierto que los inicios fueron difíciles, el avance en ciertas características, como la predicción de la pigmentación, despuntó gracias a la existencia de un reducido conjunto de genes que proporcionaban una gran cantidad de información fenotípica (101). Comentemos los distintos rasgos de interés comenzando por la predicción de color de ojos, pelo y piel continuando con otros rasgos de mayor complejidad, pero de gran interés en el campo forense.

El estudio y predicción de rasgos relacionados con la pigmentación ha estado muy interconectado a lo largo de los años, generándose en muchas ocasiones conjuntos de marcadores que tenían como objetivo interrogar simultáneamente al menos dos características físicas asociadas a este factor. La predicción de color de ojos y de color de pelo ha mantenido una relación muy cercana, por tanto, no sorprende que la publicación de los primeros modelos de predicción de estas características se agrupe en 2007 (113,114), desencadenando en la

generación de diversos modelos conjuntos y en la identificación de SNPs pertenecientes a los genes asociados a una o a las dos categorías interrogadas. Por otro lado, la predicción de color de piel presentó una mayor complejidad, estando al principio los estudios de búsqueda de marcadores, sesgados poblacionalmente, lo que provocó que la identificación de genes relacionados con el color de piel fuese más lenta y que los primeros modelos de predicción de esta categoría no surgiesen hasta 2014 (115).

Comenzando con la predicción de color de ojos, es notable la importancia que han presentado ciertos genes en el desarrollo de estos modelos. En todos los conjuntos de SNPs presentados para la predicción de esta característica se encuentran *loci* presentes en los genes *OCA2* y *HERC2*, observándose que, en la mayoría de los casos, alguna o varias posiciones asociadas a estos genes proporcionaban la mayor parte del peso predictivo de los modelos (116,117). Gracias a la identificación de marcadores asociados a este rasgo, en 2009 se generó el primer modelo que permitía realizar una clasificación entre tres categorías, color de ojos marrón, azul e intermedio (118). Liu et al. (118) desarrollaron un modelo basado en seis SNPs que proporcionaba unos valores de área bajo la curva (AUC) de 0,93, 0,91 y 0,72 para color de ojos marrón, azul e intermedio, respectivamente. Este trabajo se convirtió en uno de los pilares que permitió dos años después, en 2011, la creación del primer sistema optimizado y validado para la predicción de color de ojos de uso forense, el IrisPlex (116). Este sistema presentaba una alta sensibilidad, obteniéndose perfiles completos con hasta 30 pg de ADN de partida, y fue validado empleando muestras de las poblaciones recogidas en el HGDP-CEPH, demostrando que las predicciones del sistema no eran dependientes del origen biogeográfico (119,120).

La predicción de color de cabello, de forma un poco más lenta, siguió la estela del rasgo descrito anteriormente. Al igual que en el caso anterior, un gen, el *MC1R*, destacó notablemente, desarrollándose modelos compuestos íntegramente por *loci* pertenecientes a este gen (121). En los inicios, muchos trabajos se centraron en la predicción del fenotipo pelirrojo y no fue hasta 2011, de la mano de Branicki et al. (122) cuando se desarrolló el primer modelo completo de predicción de color de pelo. Empleando un conjunto de 22 SNPs obtuvieron valores de AUC de 0,93, 0,87, 0,82 y 0,81 para la predicción de color de pelo pelirrojo, negro, marrón y rubio, respectivamente. Este trabajo fue la puerta para que en 2013 el sistema IrisPlex evolucionase combinando el sistema previamente validado para color de ojos y este nuevo modelo de predicción de color de pelo, actualizando su nombre a HIrisPlex (120). Se siguieron los mismos pasos de validación que su predecesor aumentando a 63 pg el ADN de partida necesario para la obtención de perfiles completos (119). La evaluación de los modelos existentes con otros conjuntos de datos y la identificación de nuevos marcadores pusieron de manifiesto una posible limitación del sistema, un porcentaje de predicción correcto inferior para el grupo de color de color de pelo marrón, obteniéndose valores de AUC de 0,93, 0,86, 0,64 y 0,88 para color de cabello pelirrojo, negro, marrón y rubio, respectivamente (123). Ante estos resultados, se plantea la necesidad de la realización de más estudios que evalúen los marcadores identificados en pro de crear un modelo de predicción que presente porcentajes de clasificación elevados para todas las categorías analizadas.

Por último, el desarrollo de modelos de predicción de color de piel se vio ralentizado por el sesgo poblacional que presentaban los primeros estudios de búsqueda de marcadores. Éstos emplearon población perteneciente al mismo continente, centrándose en Europa (114) y Asia (124), cuando las mayores diferencias de color se observan a nivel continental (101). Aún con estas limitaciones, en 2014 Maroñas et al. (115), evaluando población europea y no europea, desarrollaron un modelo de predicción de color de piel compuesto por seis SNPs con el que obtuvieron un porcentaje de clasificaciones correcto de 98,3%, 92,7% y 83,7% para color de piel blanca, negra e intermedia, respectivamente. En años posteriores se continuaron identificando nuevos marcadores empleando un mayor número de poblaciones (125), lo que desembocó en una nueva actualización del sistema HIrisPlex en 2018 (62). Esta nueva versión, llamada HIrisPlex-S, se compone de 41 SNPs permitiendo predecir el color de ojos, pelo y piel simultáneamente. En el caso de pigmentación de la piel se categorizó en cinco grupos obteniéndose AUCs de 0,96, 0,88, 0,73, 0,72 y 0,74 para la predicción de color de piel de oscuro a negro, oscuro, intermedio, claro y muy claro, respectivamente.

Gracias a estos sistemas la predicción de la pigmentación se ha establecido en el campo forense abriendo paso al estudio de otras características externas visibles, generándose modelos de predicción de morfología del pelo, color de cejas, pecas, encanecimiento del pelo, calvicie masculina o peso corporal (106). Estos modelos de predicción están menos asentados y presentan valores de AUC entre 0,6 y 0,7 para alguna de las categorías analizadas. La predicción de estos rasgos aún se encuentra en un desarrollo temprano, por lo que será necesario esperar y seguir de cerca los avances en este campo a fin de comprobar su impacto y aplicabilidad rutinaria en el ámbito forense.

Al igual que en casos anteriores, la predicción de características físicas complejas se ve directamente beneficiada por los avances en secuenciación masiva en paralelo. La posibilidad de generar paneles con una gran cantidad de marcadores facilita la construcción de modelos de predicción de características complejas, como por ejemplo la morfología facial. En los últimos años gracias a estudios de asociación de genoma completo (GWAS) se han identificado multitud de *loci* genéticos asociados con el desarrollo facial (63), dando pie a que en el futuro se desarrollen modelos de predicción que permitan generar retratos moleculares a partir del ADN obtenido en la escena de un crimen. Para que esto ocurra aún falta mucho, pero es altamente probable que en los próximos años observemos un aumento en el número de publicaciones relacionadas con este tipo de estudios. Estos avances deben ir acompañados de nuevas discusiones éticas en relación a la predicción de estas características, ya que estas pruebas a fecha de 2019 solo estaban legalmente reguladas y permitidas en tres países de la Unión Europea (Alemania, Países Bajos y Eslovaquia) (126).

Aún quedan muchos aspectos que explicar del campo forense y se podría profundizar mucho más en los temas comentados, pero eso se alejaría del propósito de esta tesis. El objetivo de esta primera parte es proporcionar una visión general sobre algunos aspectos relacionados con la actualidad de la disciplina para ubicarnos en el marco espacial en el que se desarrollan los trabajos que se presentarán en las sucesivas secciones. Por tanto, a partir de este momento,

nos centraremos en introducir la temática especifica asociada a este trabajo, deteniéndonos con más calma en los aspectos que la componen y en su evolución a lo largo del tiempo.

1.4. Epigenética

El cuerpo humano es un sistema complejo que alberga muchos secretos todavía desconocidos, manteniéndose ocultos ante nuestros ojos hasta que se hace la pregunta correcta a una parte concreta del sistema. En relación con esto, la ciencia no quiere estancarse, pero como planteaba Isaac Asimov en su libro Introducción a la Ciencia, *"Es imposible captar cada detalle en un momento concreto sin quedarse rezagado inmediatamente"*. Siempre hay otro secreto, los descubrimientos generan una cascada de interés que se traduce en la generación de nuevo conocimiento en un proceso infinito. En el siglo XXI se logró descifrar el genoma humano completo, la secuencia donde se almacena toda la información genética, pero ¿cómo debemos leerla e interpretarla? Tiene que existir algún método de control que permita que en más de 200 tipos de células diferentes con el mismo código genético se expresen conjuntos de genes distintos. El código genético necesita algo que esté sobre el (epi-, procedente del griego antiguo *επί*) y que nos permita interpretarlo correctamente. El término que dará nombre a este conjunto de mecanismos se definió en 1942 de la mano de Waddington (127), la epigenética. El código genético contiene la información y el epigenoma es un decodificador que permite interpretar cuándo, dónde y cómo se debe usar esa información. Los patrones epigenéticos se definen como cambios reversibles y mitóticamente heredables, que regulan la expresión génica y que ocurren sin alterar la secuencia subyacente del ADN. Estos patrones son dinámicos ya que son el resultado de la interacción de genética, medioambiente y factores estocásticos, variando estos dos últimos a lo largo de la vida del individuo (128). De este modo, el epigenoma difiere entre poblaciones celulares del mismo organismo, a lo largo de la vida del individuo, ante ciertos desordenes de la salud y ante ciertos estilos de vida (129).

Desde su descubrimiento y teniendo en cuenta su importancia en la expresión génica, la epigenética ha estado altamente relacionada con el estudio de diversas enfermedades. A raíz de estos estudios, la influencia de los estilos de vida sobre el riesgo de padecerlas o la relación entre el epigenoma y el envejecimiento ha cobrado también gran interés. Al generarse un mayor conocimiento sobre estos mecanismos, nuevas aplicaciones fueron desentrañándose y su estudio despertó interés en diversos ámbitos, entre los que se encuentra la genética forense. Para evaluar dichas aplicaciones debemos tener en cuenta que el epigenoma está compuesto por diversas marcas, las cuales serán enumeradas a continuación, centrándonos en aquella que ha ganado mayor interés a nivel forense.

1.4.1. Marcas epigenéticas

Un todo como el epigenoma no se podría entender e interpretar si no se conocen las piezas que lo componen. La epigenética está conformada por tres capas profundamente interconectadas: modificaciones de histonas, ARNs no codificantes y metilación del ADN (130), estas marcas se recogen brevemente en el esquema que se muestra en la Figura 4. Dichas marcas son alteraciones dinámicas que tienen como objetivo modular la expresión génica y están

controladas por un conjunto de enzimas que las establecen, las eliminan o las mantienen. Estas marcas epigenéticas actúan en combinación o de forma secuencial construyendo un engranaje denominado epigenoma humano o código epigenético. Comentemos de forma breve en qué consisten estas modificaciones.

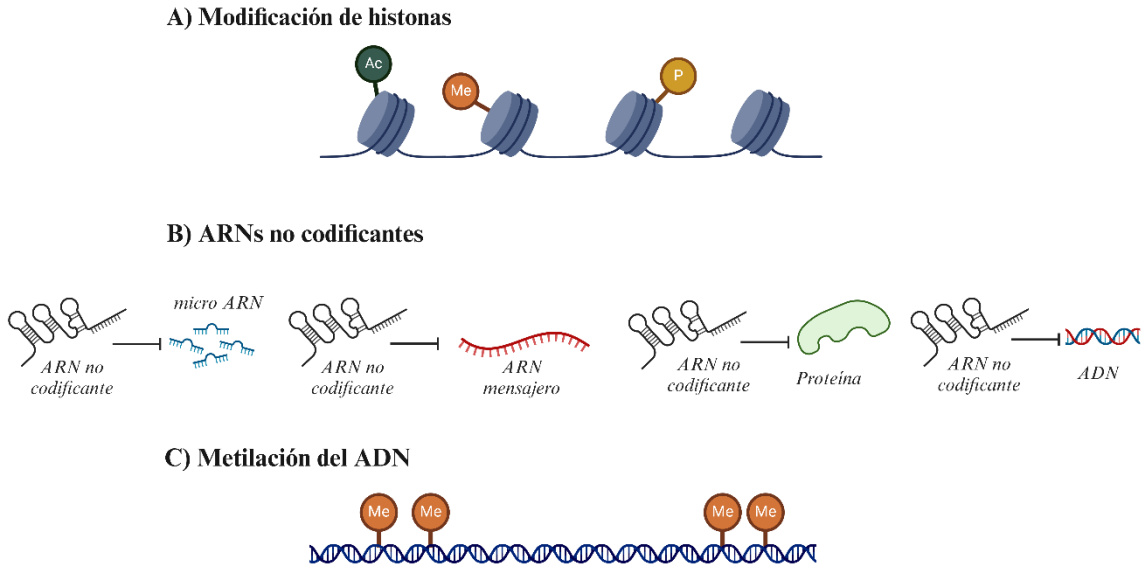


Figura 4. Esquema que representa las diferentes marcas epigenéticas que componen el epigenoma. A) Modificaciones que sufren las histonas: *Ac* se refiere a acetilación, *Me* se refiere a metilación y *P* se refiere a la fosforilación. B) Regulación génica llevada a cabo por los ARNs no codificantes interaccionando con micro ARNs, ARNs mensajeros, proteínas y ADN. C) Regulación génica ocasionada por la metilación del ADN, *Me* hace referencia a la metilación. Figura creada con BioRender.

1.4.1.1. Modificaciones de histonas

Las diferentes marcas epigenéticas que comentaremos controlan la expresión génica mediante el impedimento estérico en distintos niveles. La primera de ellas, las modificaciones de histonas, está relacionada con el empaquetamiento del ADN dentro de los cromosomas. En concreto con el nucleosoma, que se compone aproximadamente de 146 pares de bases enrollados sobre ocho histonas, conjunto que se conoce como octámero (a muchos nos sonará la asociación visual de esta estructura con un collar de cuentas). Las histonas al principio fueron consideradas inertes y no se les asociaba ninguna función aparte de su participación en el plegamiento del ADN. No fue hasta 1960 cuando Allfrey et al. (131) demostró que estas proteínas estaban asociadas a la transcripción. Gracias a esto, se identificaron aminoácidos específicos en los extremos de las histonas que pueden ser modificados mediante la adición de grupos acetilo, metilo y fosfato entre otros, como se puede observar en la Figura 4. Estas modificaciones provocan la compactación o desintegración de las estructuras nucleosomales, hecho que limita o facilita el acoplamiento de factores de transcripción al ADN. El estudio de estas modificaciones postranscripcionales de histonas se ha centrado principalmente en su relación con diversas enfermedades, siendo su impacto en el ámbito de la genética forense muy reducido o inexistente.

1.4.1.2. ARNs no codificantes

Pasamos ahora al siguiente nivel de control, en términos dimensionales, de la expresión génica. La historia de los ARNs no codificantes se remonta a 1965 cuando Robert W. Holley et al. caracterizaron, a partir de una levadura empleada en pastelerías, el primer ARN de transferencia (132). Gracias a este descubrimiento, que sería premiado con el Nobel de Fisiología o Medicina en 1968, se avanzó en la interpretación del código genético y su función en la síntesis de proteínas. El ARN no codificante (ARNncs) tiene diversas funcionalidades, pero en este caso, nos centraremos en su control sobre el proceso de traducción. Estas moléculas funcionales no se traducen a proteínas e interfieren, como se observa en la Figura 4, en la funcionalidad de otros ARNs, proteínas y el propio ADN, controlando, entre otras cosas, la expresión génica. La complementariedad de los ARNncs a ARN mensajeros es un mecanismo que regula su degradación y la parada de la traducción, generando complejos de silenciamiento. La relación de esta marca epigenética con el ámbito forense es más amplia, evaluándose su potencial en diversas aplicaciones entre las que se encuentran: la identificación de tejidos y fluidos, la causa de la muerte, estimación de la edad y la discriminación de gemelos monocigóticos (133). Aunque las moléculas de ARN han sido estudiadas en el campo forense, su inestabilidad y susceptibilidad a la degradación han dificultado, por el momento, su implementación (134).

1.4.1.3. Metilación del ADN

A partir de este punto entramos en la temática principal de esta tesis, la marca epigenética de referencia en genética forense, la metilación del ADN, representada de forma esquemática en la Figura 4. Uno podría pensar que ya eran horas después de todo lo contado anteriormente, pero parafraseando a un mago muy conocido diré "*Un escritor de tesis nunca llega tarde. Ni pronto. Llega justo cuando se lo propone*". Ha llegado el momento de profundizar en esta modificación química que ha suscitado tanto interés en diversos campos de la biología molecular.

En primer lugar, debemos preguntarnos ¿qué es la metilación del ADN? Su nombre presenta prácticamente una definición literal de lo que representa, pero esta marca epigenética tiene algunas reglas importantes. Esta modificación consiste en la adición de un grupo metilo ($-CH_3$) al carbono 5' de citosinas seguidas de guaninas (dinucleótido CpG). Es importante destacar que esta adición se produce en citosinas en organismos eucariotas, mientras que en procariotas se pueden metilar tanto citosinas como adeninas (135). Pero ¿cómo modula la expresión génica esta marca epigenética? Para ello entran en juego las denominadas islas CpG (CGIs), definiéndose como regiones de un tamaño igual o superior a 500 bp con un porcentaje de CG superior al 50%. El 70% de estas regiones ricas en dinucleótidos CpG se han identificado en promotores de genes en vertebrados, situándose el resto en regiones inter- e intragénicas (136). Al estar situadas en promotores, la expresión génica se regula en base al grado de metilación de esas regiones. Si las CGIs se encuentran hipometiladas se promueve la transcripción mientras que, cuando están hipermetiladas se produce un silenciamiento del gen (135). Esto nos lleva a preguntarnos ¿quién se encarga de metilar esas posiciones? Esta labor

la llevan a cabo las ADN metiltransferasas, enzimas que catalizan la unión de grupos metilos al átomo C-5 de la citosina. En la Figura 5 se presenta un resumen de los procesos asociados a la introducción, en el ADN, de esta modificación epigenética. Generalmente las metiltransferasas se han clasificado en dos grupos, metiltransferasas de mantenimiento (DNMT1) o de *novo* (DNMT3a y DNMT3b), pero es más flexible de lo que parece. Las enzimas DNMT1 principalmente se ocupan de copiar los patrones de metilación durante la replicación, pero también son capaces de llevar a cabo metilación de *novo*. Por otro lado, DNMT3a y DNMT3b son las principales encargadas de establecer los patrones de metilación durante el desarrollo temprano, con capacidad de mantenimiento apoyando, de ser necesario, a la DNMT1 durante la replicación (137). Pero, como nos enseñó Kelsier (en el libro El Imperio Final de Brandon Sanderson) "*Es cosa de la naturaleza. A cada empujón le corresponde un tirón. Una consecuencia.*", por tanto, existe un mecanismo contrario que regula la eliminación de esta marca epigenética. Este proceso, la desmetilación del ADN, puede producirse, como se muestra en la Figura 5 en color rojo, de forma pasiva o activa. La desmetilación pasiva, identificada por una línea descontinuada en la Figura 5, ocurre durante la replicación y se debe a la inacción de la proteína DNMT1, no manteniendo la metilación en las nuevas cadenas de ADN sintetizadas tras la división celular. Actualmente existen procesos farmacológicos que permiten controlar la desmetilación pasiva y se emplean para activar genes supresores de tumores silenciados por metilación (138). Por otro lado, tenemos la desmetilación activa que es independiente de la replicación del ADN y que está mediada por procesos de oxidación llevados a cabo por proteínas de la familia TET. Esta desmetilación se produce mediante la oxidación de 5-metilcitosina (5-mC) en 5-hidroxilmetilcitosina (5-hmC), el proceso continúa transformándola en 5-formilcitosina (5-fC) y, por último, en 5-carboxilcitosina (5-caC). Estos productos de oxidación presentes en el ADN se reparan introduciendo una citosina no modificada por escisión, proceso llevado a cabo por la Timina ADN Glicosilasa (TDG) mediada por la ruta de reparación por escisión de bases (BER) (139). Es importante destacar, fijándonos en la Figura 5, que la escisión e introducción de una nueva citosina puede producirse en dos etapas de la desmetilación, en presencia de la 5-fC o de la 5-caC. Ahora que disponemos de una visión general del funcionamiento de la metilación y desmetilación del ADN en nuestro organismo, podemos evaluar el impacto generado por esta marca epigenética.

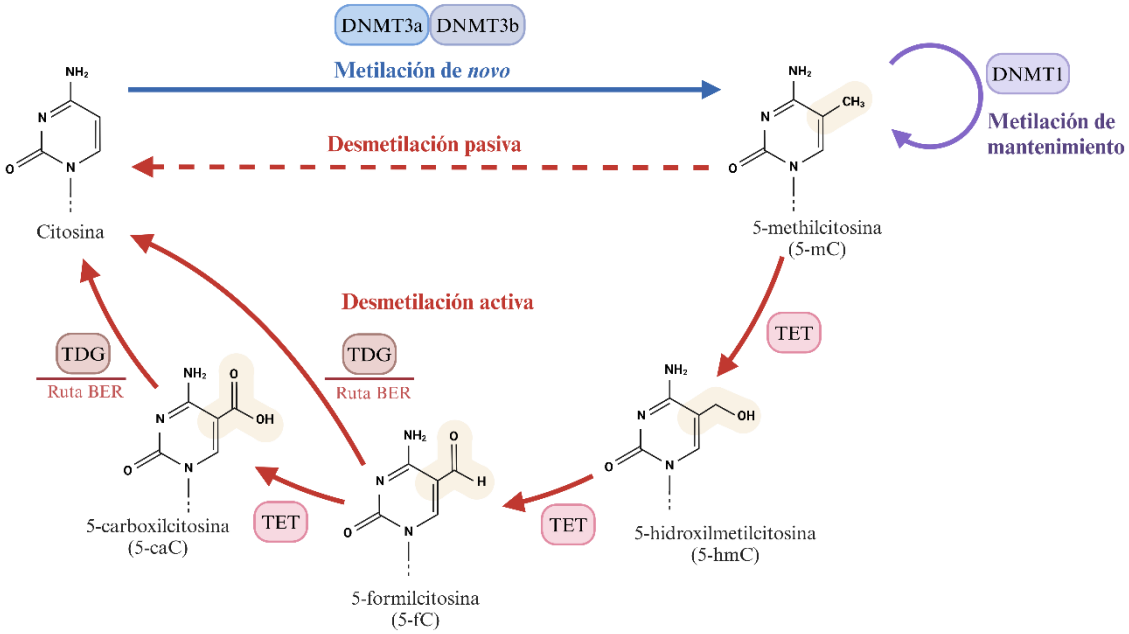


Figura 5. Esquema del proceso de metilación y desmetilación que sufren las posiciones CpG del ADN. TET hace referencia proteínas de dicha familia, TDG a la timina ADN glicosilasa y BER a la ruta de reparación por escisión de bases. Figura creada con BioRender.

La metilación del ADN es la marca epigenética que mayor interés ha suscitado, generándose una gran cantidad de conocimiento en relación con sus capacidades. Todos estos estudios han definido, en muchos casos, las posibles aplicaciones de este biomarcador. Su importancia en la regulación génica incentivó su estudio clínico, publicándose por primera vez en 1983 su asociación con el cáncer (140). Este primer trabajo fue el pistoletazo de salida que propiciaría su análisis en múltiples enfermedades. El estudio de esta capacidad de regulación llevó a la identificación de diferentes patrones de metilación tejido específicos, siendo no sólo importante el *locus* seleccionado, sino también el tejido analizado (141). Al definirse como una marca dinámica se observó que gemelos monocigóticos presentaban divergencias en su epigenoma producidas por la edad (129) y por enfermedades (142). Estos estudios propiciaron la capacidad de identificar diferencias entre dos personas con genoma idéntico y los patrones de metilación asociados al envejecimiento. Manteniendo la atención en el dinamismo de este biomarcador, se centró el foco en el efecto de procesos activos que puedan tener influencia en algunas enfermedades. Con esto en mente, la influencia de los estilos de vida sobre ciertas dolencias y su efecto en los patrones de metilación fueron evaluados (143). El estudio de los patrones de metilación relacionados con el estilo de vida, el consumo de drogas (tabaco, alcohol y drogas duras) y el ejercicio físico ganaron protagonismo aportando contexto a enfermedades complejas, como por ejemplo el cáncer.

Como se puede observar, el ámbito clínico ha impulsado el interés y el avance del conocimiento epigenético y la onda expansiva llegó hace años al campo forense. Gracias a todos estos estudios, la comunidad forense pudo identificar el gran potencial de la metilación del ADN adaptando las aplicaciones anteriormente citadas. Antes de comentar en profundidad dichas aplicaciones en el ámbito que nos compete, es importante que primero establezcamos un

pilar importante para el establecimiento de este biomarcador como la marca epigenética de preferencia. El análisis de la metilación del ADN es fácilmente abordable a nivel técnico, condición de gran importancia en genética forense, comentemos, por tanto, las técnicas más importantes empleadas en el análisis de la metilación.

1.4.2 Técnicas de detección de la metilación del ADN

Ahora que sabemos definir la marca epigenética que estamos estudiando y antes de profundizar en sus aplicaciones, debemos establecer los procesos que nos permiten su análisis. A lo largo de este apartado se llevará a cabo una explicación breve de las cuatro técnicas más usadas para el estudio de la metilación del ADN en el ámbito forense. Es lógico preguntarnos ¿cómo podemos diferenciar entre una citosina sin metilar y una metilada? Estamos acostumbrados a estudiar la secuencia del ADN identificando adiciones, deleciones o sustituciones de bases nitrogenadas enteras, pero ¿una modificación sobre una de ellas? Para afrontar este estudio, la técnica de mayor relevancia y la más empleada es la conversión con bisulfito. Este pretratamiento del ADN nos permite identificar esta modificación generando una diferencia entre dos bases idénticas a nivel de secuencia, pero diferentes a nivel epigenético.

1.4.2.1 Conversión con bisulfito sódico

El proceso de conversión de ADN con bisulfito sódico fue descubierto en 1970 (144,145) y se empleó en 1992 (146) para desarrollar el análisis de la metilación del ADN. Esta reacción química, esquematizada en la Figura 6, se basa en la capacidad del bisulfito sódico de selectivamente desaminar citosinas no metiladas, estando las metiladas protegidas de este tratamiento gracias al grupo metilo. La desaminación que sufren las citosinas no metiladas conlleva a una transformación de estas bases en uracilo, el cual, al no formar parte de las bases primarias del ADN, durante la PCR se amplifica como una timina. Este proceso nos permite estudiar la metilación del ADN en base a la detección de dos nucleótidos, representando uno la cantidad de citosinas metiladas y el otro la cantidad de citosinas no metiladas. ¿Cómo se traduce esto a nivel de secuencia? Al analizar muestras tratadas con bisulfito se evalúa la relación entre nucleótidos C/T (amplificación en sentido, cadena +) o la relación entre G/A (amplificación en antisentido, cadena -) de una o varias posiciones de interés como si fuesen SNPs. Si bien es cierto que la reacción química llevada a cabo por el bisulfito es muy agresiva, provocando la degradación del ADN en el proceso, se debe destacar que gracias a los *kits* comerciales desarrollados actualmente el proceso se puede realizar con gran efectividad y alto rendimiento (147). Esta optimización es una ventaja fundamental para el ámbito forense permitiendo el uso de una menor cantidad de ADN de partida.

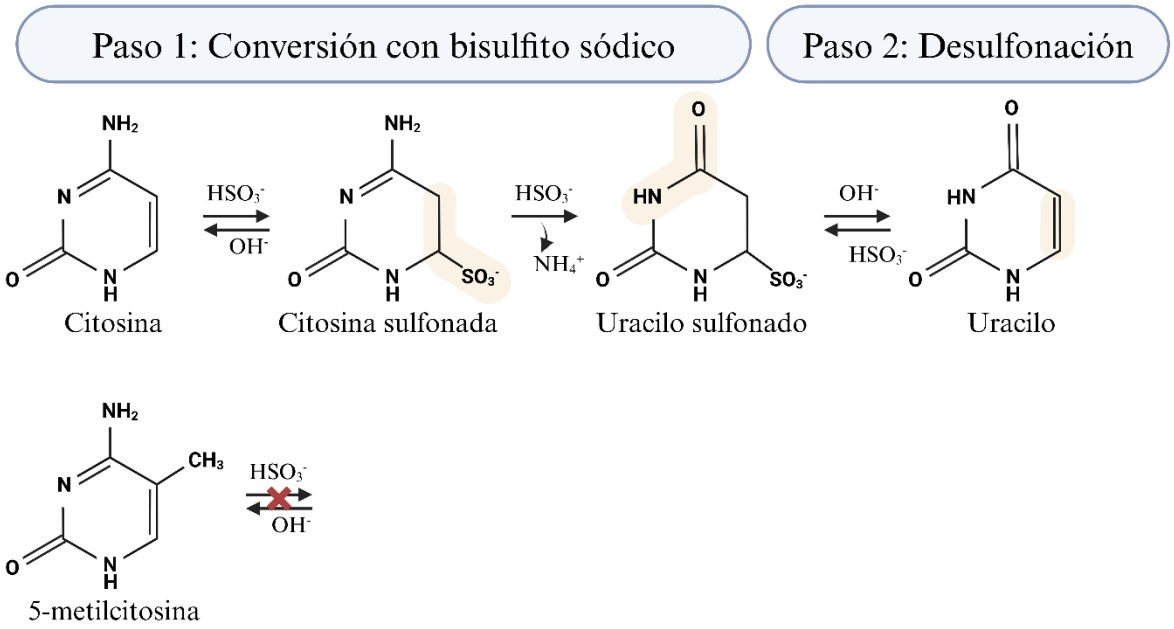


Figura 6. Esquema del proceso de conversión con bisulfito sódico de las citosinas no metiladas. Figura creada con BioRender.

El análisis de ADN convertido consiste en el empleo de metodologías cuantitativas o semicuantitativas que nos permitan definir la cantidad de citosinas metiladas y no metiladas en la/las posiciones de interés. A partir de esta información se calculará el nivel de metilación. Existen diversas metodologías que permiten analizar la metilación del ADN, las cuales presentan distintos alcances en base a los objetivos del estudio, pero, en este caso, nos centraremos en aquellas que tienen o han tenido una mayor repercusión en el campo forense.

1.4.2.2 Metodologías de descubrimiento

El descubrimiento de posiciones CpG y la obtención de sus patrones de metilación es un paso crucial en el desarrollo de las diversas aplicaciones asociadas a este biomarcador. En los inicios, al igual que en el caso del estudio de la secuencia genómica, se emplearon tecnologías que buscaban evaluar simultáneamente la metilación del ADN a lo largo de todo el genoma. Dichas metodologías presentaban una ligera adaptación, como veremos en el resto de las técnicas empleadas para el estudio de este biomarcador, en forma de pretratamiento de ADN, la conversión con bisulfito sódico, denominándose secuenciación con bisulfito de genoma completo (WGBS). Gracias a los estudios realizados con dicha metodología se reforzaron o se identificaron características asociadas a este biomarcador, como por ejemplo una disminución de la densidad de CpGs metiladas en zonas del genoma de interacción con proteínas o la existencia de posiciones no CpG metiladas, observadas, en una proporción reducida (10%-40%), en células madre embrionarias (148–150). Pero, como dijo Sempronio en La Celestina, obra atribuida a Fernando de Rojas, "*Contentémonos con lo razonable, no sea que por querer más lo perdamos todo, que quien mucho abarca, poco aprieta*.", este análisis que intenta obtener simultáneamente tanta información presentó limitaciones derivadas de ese mismo objetivo. Esta metodología, aunque potente, se enfrentaba a ADN convertido mediante un

proceso agresivo con el mismo y buscaba identificar las, aproximadamente, 28 millones de posiciones CpG presentes en el genoma. Esta magnitud se tradujo en una baja resolución (151), sesgo específico de secuencia (152), la necesidad de reducción del análisis a fracciones más pequeñas del genoma (153) o una ineficiencia general proporcionando el 70%-80% de las lecturas poca información en relación con la metilación del ADN (154). Ante estos obstáculos se generaron alternativas, reduciendo en gran medida el número de variables interrogadas (155), originándose entre ellas los ampliamente utilizados paneles *BeadChip* de Illumina.

Los paneles de Illumina, que también requieren el mismo pretratamiento de ADN que los WGBS, emplean sondas complementarias a las regiones de interés, amplificando un número de posiciones más limitado. Esta reducción los convierte en análisis más eficientes en lo relativo a su coste y tiempo de análisis, así como les proporciona una mejor especificidad y una mayor resolución. Sería en 2009 cuando el primero de estos paneles vería la luz, el *Infinium HumanMethylation27 BeadChip* (27K), comprendiendo un total de 27.579 posiciones CpG (156). Aunque este número parezca muy elevado, solamente representa, aproximadamente, el 0,098% del metiloma (≈28 millones de CpGs), pero esto solo era el inicio. Teniendo en cuenta esta metodología, desarrollaron sondas para 16 millones de posiciones CpG que emplearon pocos años después para ir incrementado la capacidad de su tecnología *BeadChip*. Con esto en mente, en 2011, se publicó el siguiente panel denominado *Infinium HumanMethylation450 BeadChip* (450K), comprendiendo 485.577 posiciones CpG (con más del 90% de las posiciones presentes en el 27K) (157). La publicación de este panel y el auge del estudio de la metilación convertiría a este *chip* en la plataforma de elección para el descubrimiento y selección de posiciones de interés, contribuyendo a la construcción de múltiples modelos de interés forense. El salto con respecto a su predecesor, el 27K, es sustancial, pero solamente representa un 1,7% de las posiciones CpG repartidas por nuestro genoma por lo que el avance debía continuar. Cinco años después llegaría la siguiente versión, comprendiendo 853.307 CpGs (91,1% de las presentes en el 450K) denominado *MethylationEPIC BeadChip Infinium* (EPIC) (158). La diferencia de esta nueva versión con respecto a las anteriores, que centraban sus esfuerzos, predominantemente, en la identificación de posiciones presentes en islas CpG y promotores, es la apuesta por marcadores en regiones regulatorias, como potenciadores o sitios de unión de factores de transcripción, y la detección de 5'-hidroximetilcitosinas, nuevas marcas epigenéticas asociadas con la regulación génica (159). Este panel recibiría recientemente una actualización, *Infinium MethylationEPIC v2 BeadChip* (EPIC v2), ampliando su conjunto de posiciones a 937.690 posiciones CpG (160). Este acercamiento segmentado ralentiza el proceso, abarcando por el momento el 3,35% del metiloma, pero se obtiene una mayor resolución permitiendo el desarrollo y obtención de conocimiento más estable.

Este conjunto de paneles ha propiciado el avance de los estudios de metilación del ADN, proporcionando un método de descubrimiento asequible y abordable, tanto bioinformática como metodológicamente. A nivel forense han proporcionado, gracias a la disponibilidad de conjuntos de datos públicos, la posibilidad de evaluar la correlación de estos marcadores con diferentes factores, como por ejemplo la edad y los estilos de vida. Esta identificación ha

permitido seleccionar marcadores de interés para su validación empleando las metodologías descritas a continuación.

1.4.2.3 Espectrometría de masas

La espectrometría de masas lleva con nosotros casi 150 años, siendo su creación una sucesión de adaptaciones técnicas que culminan en lo que hoy conocemos por ese nombre. En cambio, la configuración MALDI-TOF y su aplicación en el análisis de la secuencia del ADN es algo más reciente, siendo creada en 1988 (161,162) y aplicada con este fin a principios de los años 90 (163). La espectrometría de masas tiene diferentes configuraciones que vienen definidas por las necesidades del estudio que se lleva a cabo. Para el análisis de la metilación del ADN, se define en 2005 una adaptación de la técnica bajo el nombre de EpiTYPER (164), un sistema de análisis con tratamiento por bisulfito sódico previo a la espectrometría de masas MALDI-TOF. Llevemos brevemente la explicación a un nivel un poco más técnico, evaluando las partes que componen esta tecnología. La espectrometría de masas se basa en la obtención de iones en estado gaseoso de las moléculas de interés y su posterior separación en base a su relación masa/carga. Un espectrómetro de masas lo conforman tres elementos principales: una fuente de ionización, un analizador de masas y un detector. En el caso de EpiTYPER la fuente de ionización es MALDI (desorción/ionización láser asistida por matriz), técnica de ionización suave que, empleando una matriz orgánica, permite el análisis de biomoléculas y moléculas orgánicas grandes. Una vez convertidas las moléculas de interés en iones se dirigen al analizador de masas, TOF (tiempo de vuelo). En este analizador, los iones son acelerados adquiriendo una elevada energía cinética propiciando que, idealmente, todos los iones tengan la misma energía, variando sus velocidades solamente según su ratio masa/carga. En el caso del sistema MALDI-TOF la ratio masa/carga suele ser equivalente a la masa molecular del analito, observándose un pico correspondiente a dicha masa cuando llega al detector. Este proceso, visto así, puede ser difícil de extrapolar a nuestro caso, por tanto, ¿cómo se aplica en el análisis de metilación del ADN? Para seguir este proceso con mayor facilidad podemos recurrir a la Figura 7, donde se esquematizan los pasos seguidos para analizar este biomarcador con espectrometría de masas. En primer lugar, se convierte con bisulfito sódico el ADN genómico (CG posiciones metiladas, UG posiciones no metiladas) y se amplifican las regiones de interés (CG posiciones metiladas, TG posiciones no metiladas). A continuación, se transcribe *in vitro* a ARN (GC posiciones metiladas, AG posiciones no metiladas) y se escinde en fragmentos definidos por la posición de los uracilos presentes en la secuencia empleando una endoribonucleasa. Al analizar estos fragmentos con un equipo de espectrometría de masas MALDI-TOF se observan diferencias de tamaño de 16 daltons por cada CpG metilada, valor que hace referencia a la diferencia de masa molecular que presenta la guanina (posición metilada) con respecto a la adenina (posición no metilada).

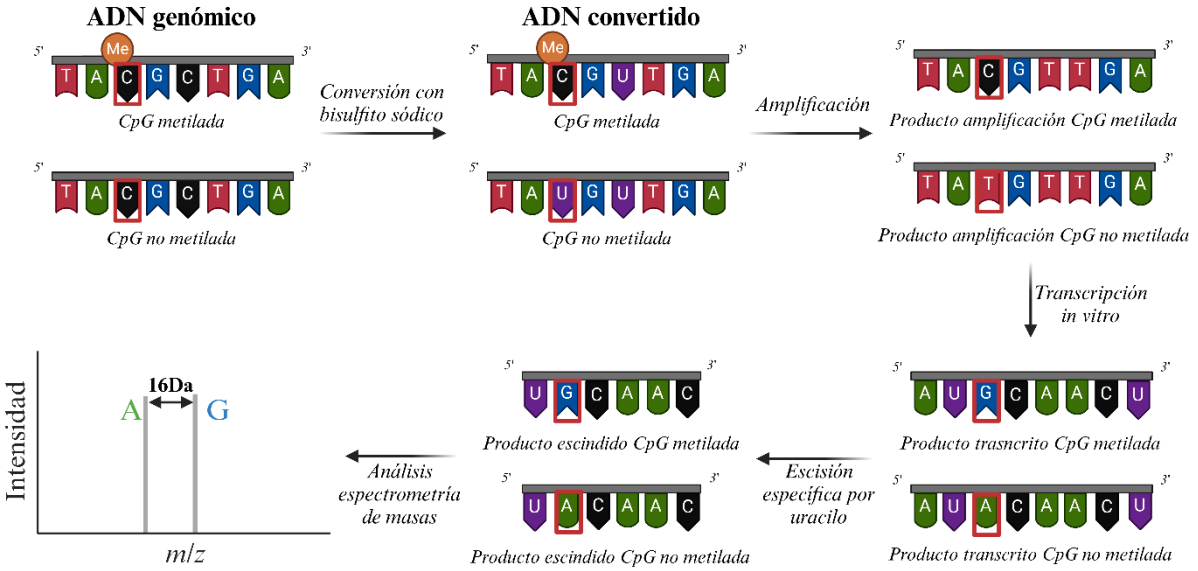


Figura 7. Esquema del proceso de análisis de metilación del ADN mediante espectrometría de masas. La relación masa/carga se representa como m/z. La citosina analizada en este esquema se resalta con un recuadro rojo. Figura creada con BioRender.

La tecnología EpiTYPER permite analizar cuantitativamente la metilación del ADN en regiones genómicas de 100-600 bp, siendo un proceso altamente automatizado y útil para analizar un gran número de muestras o regiones en un sólo análisis. Estas ventajas la convirtieron en una herramienta adecuada para el descubrimiento y validación (165), sin embargo, su incapacidad para el análisis en *multiplex* la convierte en una técnica de difícil aplicación directa en casos forenses. En los últimos años, otras tecnologías que comentaremos a continuación han ganado un mayor protagonismo, siendo escasos actualmente los estudios que emplean esta metodología.

1.4.2.4 Pirosecuenciación

La pirosecuenciación es una técnica de secuenciación por síntesis que fue inventada en 1988 (166), aunque tardaría 12 años en comercializarse e implementarse por completo. El desarrollo de esta tecnología no se podría haber llevado a cabo sin el trabajo de Pål Nyrén en 1987 (167), demostrando que la polimerización del ADN podía ser monitorizada mediante la medición de la producción de fosfatasa. Esta tecnología sería rápidamente adaptada al análisis de la metilación del ADN, publicándose el primer artículo en 2007 (168). Veamos, de forma general, cómo funciona esta metodología y cómo se adecúa a este objetivo concreto, para un seguimiento más sencillo la Figura 8 esquematiza este proceso. Para poder llevar a cabo el análisis mediante pirosecuenciación se realiza, en primer lugar, una amplificación de la región de interés, del ADN convertido con bisulfito sódico, con un cebador antisentido biotinilado. Este grupo biotina permite la inmovilización del producto de PCR sobre partículas magnéticas con estreptavidina, formándose un enlace avidina-biotina. Estas esferas pueden fijarse a la pared de un tubo mediante la aplicación de un campo magnético, lo que permite la purificación del producto de PCR y los posteriores lavados. Una vez tenemos el fragmento de interés bloqueado se produce un proceso de amplificación nucleótido a nucleótido, añadiéndose cada base

nitrogenada individualmente y de forma secuencial. Cuando se produce una correcta hibridación del nucleótido complementario se libera un grupo pirofosfato que se convierte en adenosín trifosfato y oxida luciferina emitiendo luz. Esta emisión de luz generada es proporcional a la cantidad de pirofosfato producida, la cual es directamente proporcional al número de nucleótidos hibridados durante la amplificación de la secuencia de interés.

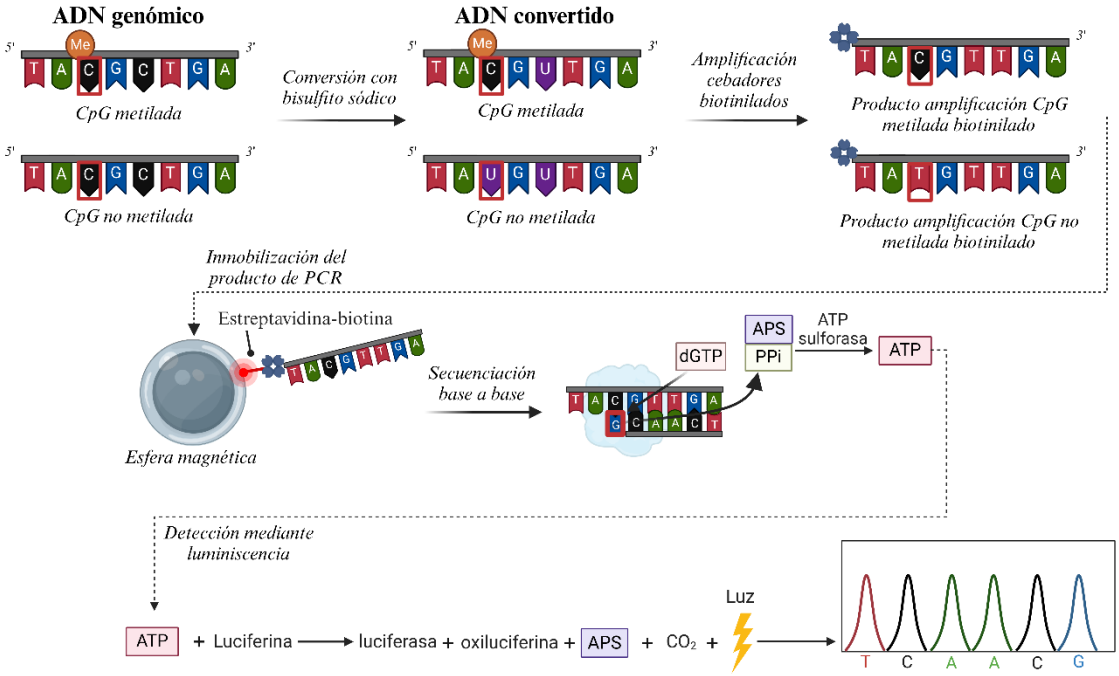


Figura 8. Esquema del proceso de análisis de metilación del ADN mediante pirosecuenciación. La citosina analizada en este esquema se resalta con un recuadro rojo. PPi hace referencia a pirofosfato, APS a persulfato amónico, ATP a adenosín trifosfato y CO_2 a dióxido de carbono. Figura creada con BioRender.

La pirosecuenciación ha sido una metodología con un gran impacto en el análisis de la metilación del ADN, desarrollándose diversos modelos de predicción empleando esta técnica. Las ventajas que presenta esta técnica son la selección personalizada de CpGs, así como la obtención de valores de metilación muy exactos. Pero al igual que la metodología anterior, existe una desventaja que limita su aplicación directa al ámbito forense (165), viéndose superada en los últimos años por técnicas que permiten el análisis en *multiplex*.

1.4.2.5 Minisecuenciación

La técnica de minisecuenciación se basa en el análisis de la secuencia de ADN mediante extensión de un único nucleótido (SBE) y fue definida en 1990 (169), asociando su aplicación para uso clínico. Sería 10 años después cuando surgiría el *kit* SNaPshot™ y su aplicación en *multiplex* (170), convirtiéndose con el paso del tiempo en la herramienta estándar para el análisis de SNPs y una metodología de interés para el análisis de la metilación del ADN. La aplicación de esta metodología al estudio de este biomarcador se planteó, a nivel clínico, en 2002 (171) pero habría que esperar hasta 2015 para que el campo forense lo adaptase y se publicase el primer modelo preliminar para predicción de la edad basado en este biomarcador empleando SNaPshot™ (172). Veamos pues cómo funciona la minisecuenciación,

esquematizada en la Figura 9, y cómo se aplica este procedimiento al análisis de la metilación del ADN. Esta metodología comienza con una amplificación de una pequeña región en la que se encuentra el SNP de interés. El producto de PCR se purifica y se realiza una segunda amplificación empleando cebadores SBE que hibridan justo antes del nucleótido de interés. Pero esta nueva PCR tiene una diferencia importante, emplea dideoxinucleótidos (ddNTPs) marcados fluorescentemente (cada nucleótido está asociado a un fluorocromo diferente) que actúan como terminadores ya que carecen de grupo hidroxilo en el carbono 3'. Tras esta minisecuenciación, el producto se purifica eliminando los ddNTPs marcados sobrantes y se lleva a cabo el análisis de las muestras mediante electroforesis capilar. Este procedimiento es fácilmente adaptable al análisis de la metilación del ADN, mediante pretratamiento con bisulfito sódico que nos permite evaluar las CpGs de interés como si de un SNP se tratase.

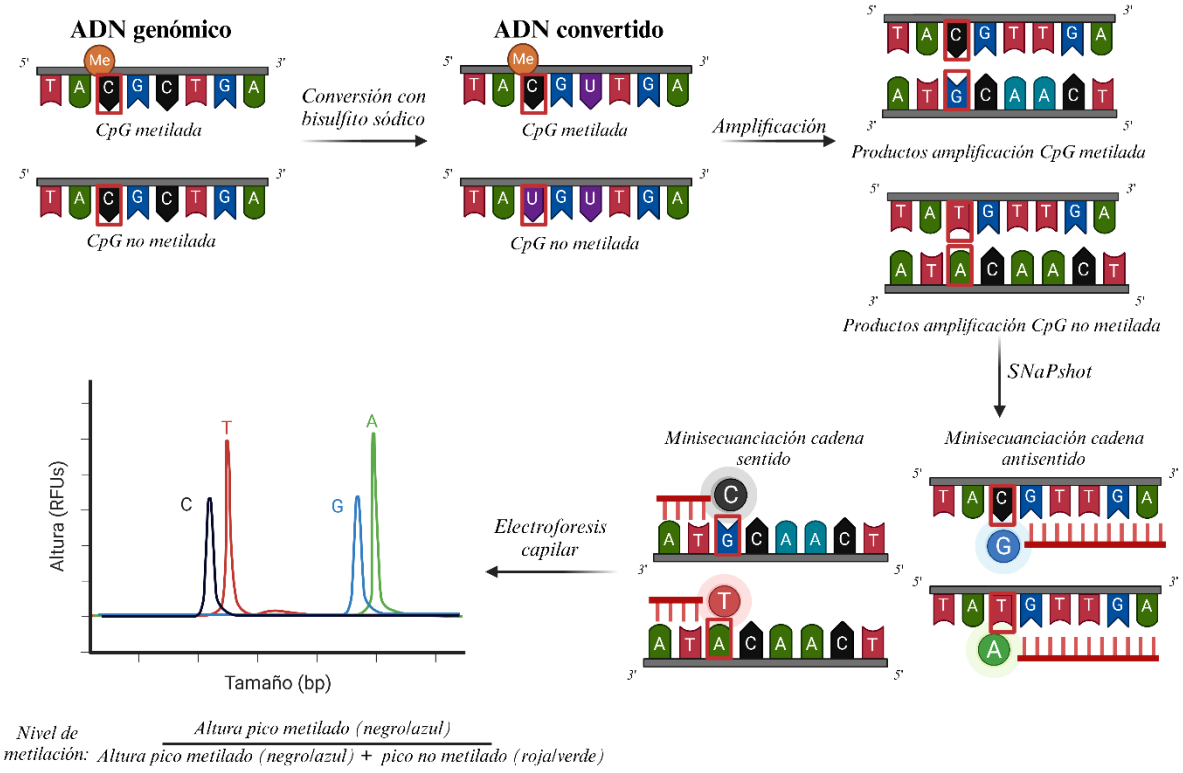


Figura 9. Esquema del proceso del análisis de la metilación del ADN mediante minisecuenciación con SNaPshot™. Marcada en rojo se presenta, a lo largo de todo el proceso, la citosina analizada en este ejemplo. *Me* hace referencia al grupo metilo presente en una citosina metilada. Figura creada con BioRender.

El uso de la minisecuenciación ha cobrado una gran importancia en los últimos años en el ámbito forense. Esta escalada se debe a sus ventajas con respecto a las técnicas comentadas anteriormente. SNaPshot™ proporciona un método sencillo que permite analizar simultáneamente múltiples posiciones, diseñar amplicones más pequeños y emplear una menor cantidad de ADN genómico de partida, proporcionando una gran ventaja frente a muestras degradas (165). Sin embargo, esta metodología tiene como desventaja la forma en la que se evalúa la metilación del ADN. Para calcular los niveles de metilación, se estudia la relación entre las alturas de los picos. Al ser detectados estos mediante fluorocromos distintos, su altura

puede verse afectada por las diferentes intensidades de señal que presentan los fluoróforos empleados.

Por último, avancemos hacia el futuro metodológico comentando una tecnología que está ganando interés y empieza a posicionarse como el protagonista en los avances de este campo.

1.4.2.6 Secuenciación masiva en paralelo

Cuando hacemos referencia a secuenciación masiva en paralelo no podemos pensar en una única técnica como en los casos anteriores. Este término hace referencia a diversas tecnologías que comparten una característica metodológica, la secuenciación masiva de moléculas de ADN clónicas que están separadas espacialmente en una celda de flujo. El lanzamiento comercial del primer sistema de MPS, Roche 454 GS-20, se produce en 2005 combinando la PCR en emulsión y la pirosecuenciación (173). El potencial de estas nuevas tecnologías no pasó desapercibido y quedó patente tres años después, cuando el heredero del instrumento anterior, el GS FLX, lograse secuenciar un genoma humano en el 3% del tiempo y con un presupuesto del 0,56% del empleado en el proyecto Genoma Humano, realizado mediante electroforesis capilar (174). La capacidad de estas tecnologías atrajo el interés del campo forense y su uso empezó a popularizarse, centrándose en la actualidad en tres tecnologías: el sistema Ion S5 de ThermoFisher, el MiSeq® FGx de Illumina y el MinION de Oxford Nanopore. Estos instrumentos son todos considerados MPS, pero ¿cuáles son sus diferencias metodológicas? Con el objetivo de explicar las similitudes y diferencias de estas tecnologías se ha diseñado la Figura 10, que permite seguir las explicaciones presentadas a continuación de una forma simple y visual. El sistema de ThermoFisher, comercializado en 2015, basa su secuenciación en los protones que se liberan durante la polimerización del ADN, detectando el cambio de pH que se produce al incorporarse un nucleótido. Veamos los pasos de forma general siendo el primer paso la amplificación de las regiones de interés. Estos fragmentos de ADN se incorporan a unas esferas para ser amplificados clónicamente mediante PCR en emulsión. A continuación, esas esferas son adicionadas a micropocillos presentes en un *chip* sobre el que se añade, de forma secuencial, soluciones que contienen, de forma individual, cada uno de los nucleótidos presentes en el ADN. Si se produce la incorporación a la cadena de ADN de la base presente en la solución, se genera un enlace covalente que libera un pirofosfato y un protón que es detectado por la base del pocillo al actuar este como un peachímetro. En el caso de Illumina, debemos destacar que fue la primera compañía en desarrollar y validar, en 2014, un equipo de MPS centrado en genética forense. En cuanto a su metodología, la primera diferencia destacable que encontramos en comparación con la anterior es la ausencia de PCR en emulsión. En el caso de los dispositivos de Illumina, los fragmentos de ADN de interés se hibridan en una celda de flujo y se produce una amplificación puente que crea grupos de moléculas de ADN clónicas. A continuación, se produce una secuenciación por síntesis mediante la adición y detección óptica secuencial de nucleótidos marcados fluorescentemente. Por último, el sistema más novedoso, aplicado al ámbito forense por primera vez en 2017, es el MinION, que basa su tecnología en la detección de cambios en la corriente eléctrica. Este dispositivo contiene una membrana con nanoporos a través de los cuales se hace pasar la secuencia de ADN de interés, identificando

los nucleótidos que lo componen al detectar cambios de corriente que se producen durante el paso por la membrana. La adaptación de estas metodologías al análisis de la metilación del ADN técnicamente es simple, para Ion S5 y MiSeq® FGx, representados en la sección uno y dos, respectivamente, de la Figura 10, sólo es necesario introducir el pretratamiento del ADN con bisulfito sódico. En el caso de MinION, sección tres de la Figura 10, dicho paso no es necesario y la lectura de cambio de corriente eléctrica puede diferenciar una citosina metilada y no metilada, pero se necesita un algoritmo entrenado para identificar esa variación de corriente correctamente (175).

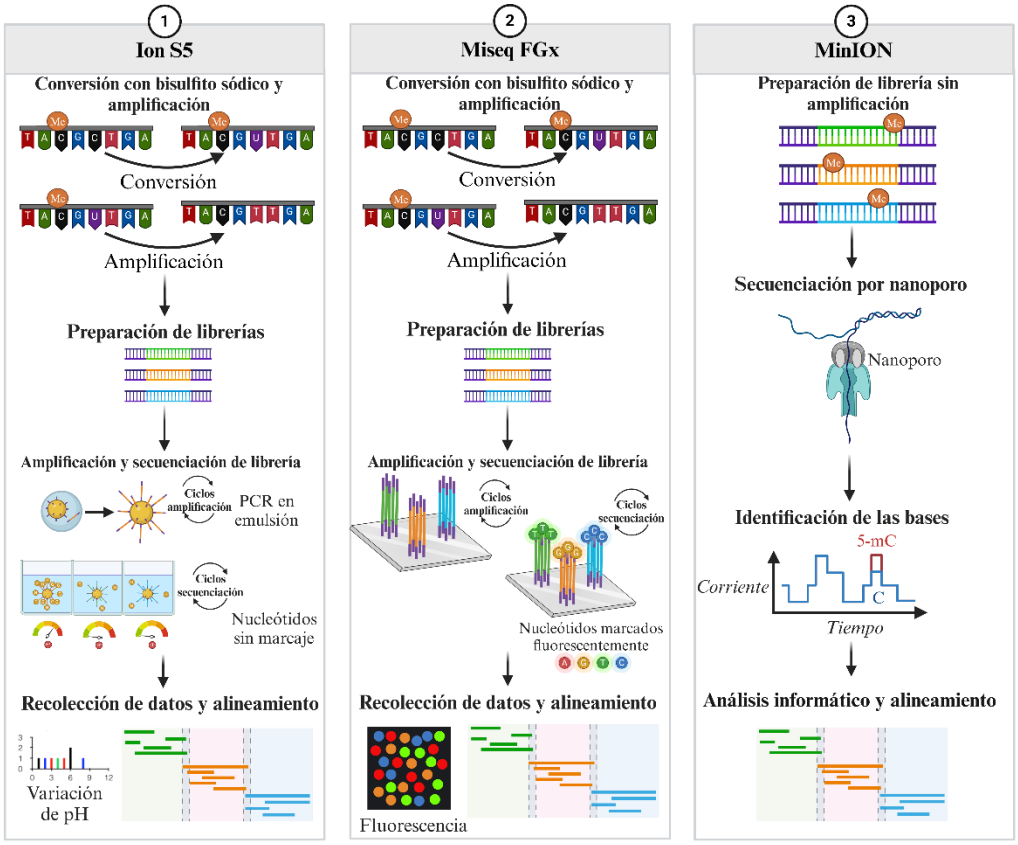


Figura 10. Esquema del proceso de análisis de metilación del ADN en las plataformas de MPS de uso forense. El primer esquema hace referencia al secuenciador Ion S5 de ThermoFisher, el segundo al MiSeq® FGx de Illumina y el tercero al MinION de Oxford Nanopore. Figura creada con BioRender.

Las ventajas que proporcionan estas metodologías han incrementado su popularidad, siendo cada vez más habitual su uso en el desarrollo de modelos forenses, como veremos en los siguientes apartados. Dichas tecnologías, en comparación con las anteriores, permiten analizar simultáneamente un número sustancialmente más elevado de marcadores generando datos de alta resolución, cualidades altamente demandadas en el ámbito forense. Lo único que ralentiza su implementación, aunque avanza rápidamente, es el coste asociado de los equipos y análisis en un contexto forense, donde es posible que no se puedan acumular un número mínimo de muestras que hagan rentable la carrera de secuenciación.

Con un conocimiento general de la metilación del ADN y de las metodologías empleadas para su análisis, podemos centrarnos en las aplicaciones forenses y los avances logrados durante los últimos años.

1.4.3 Aplicaciones forenses de la epigenética

En el ámbito forense, hacer referencia a la epigenética suele estar asociado casi inevitablemente al estudio de la metilación del ADN. Este biomarcador ha generado un merecido interés dentro de la comunidad científica de este campo gracias a los descubrimientos aportados por todos los conocimientos generados en los últimos años. Esta galaxia de publicaciones ha permitido concretar su aplicabilidad al ámbito forense, siendo considerada, como se mencionó previamente, parte de las *DNA intelligence tools* de la genética forense moderna. Su aplicación en este y otros campos ha orbitado en relación a dos características concretas: la metilación del ADN es tejido especifica y además es una marca dinámica influenciada por el ambiente. Estos factores han permitido, en el ámbito forense, definir las aplicaciones de este biomarcador a la hora de aportar información sobre el donante de una muestra de ADN. Por tanto, a continuación, se comentarán las aplicaciones más destacables de la metilación del ADN en genética forense. Si tenemos en cuenta que esta marca epigenética se ve influenciada por el ambiente y no modifica la secuencia subyacente, el estudio del epigenoma de gemelos monocigóticos, genéticamente idénticos, genera un gran interés, formándose así la idea de una primera aplicación forense.

1.4.3.1 Discriminación de gemelos monocigóticos

Recapitulando un poco lo comentado anteriormente, la genética forense basa sus análisis rutinarios en la variación genética humana, permitiendo la identificación de personas mediante el estudio de, por ejemplo, los STRs. Pero este tipo de marcadores tienen un talón de Aquiles, un escenario en el que su gran poder de discriminación se ve inutilizado. ¿Qué pasa si los genomas de dos personas son “idénticos”? Esta situación se vive con gemelos monocigóticos (MZ), que al derivar del mismo cigoto comparten una secuencia genética “idéntica”, lo que los hace indistinguibles a los ojos de los STRs y otros marcadores de uso forense. Esta situación, aun siendo poco común, presenta un problema a nivel judicial, convirtiendo las pruebas de ADN en irrelevantes. La mera existencia de estos casos y los problemas derivados de su resolución, crean la necesidad de desarrollar un método molecular que permita diferenciar a gemelos monocigóticos. Aún con sus limitaciones, la variación genética no se dio por perdida por completo, planteándose la posibilidad de emplear tecnologías de secuenciación de genoma completo en la diferenciación de gemelos MZ (176). Este acercamiento se basa en la ocurrencia de mutaciones somáticas en sólo uno de los gemelos durante el desarrollo temprano, pero presenta desventajas que limitan en gran medida su aplicación en el ámbito forense. Las mutaciones somáticas pueden producirse o no y, de existir, pueden no estar presentes en todos los tipos celulares. Aunque existen nuevas publicaciones (177) y metodologías que podrían convertir estos análisis en económicamente asequibles, sus desventajas seguirían limitando en gran medida su estandarización en rutina forense. Si la estabilidad del genoma nos limita a la

hora de diferenciar entre gemelos monocigóticos, ¿dónde podemos buscar esas diferencias? Para encontrar una respuesta a esta pregunta debemos ser un poco más flexibles, dinámicos, como el biomarcador que puede tener la solución. Al abrirse nuevos horizontes en la genética forense surge el epigenoma y los estudios de la metilación del ADN, una marca que no solo se ve definida por la genética, sino también por factores estocásticos y por la influencia del ambiente. Los gemelos monocigóticos se convierten en un modelo de gran interés a nivel clínico, en 2005 surgen los primeros estudios que buscan evaluar las diferencias entre ellos con el objetivo de definir los mecanismos de discordancia fenotípica y los efectos del epigenoma derivados de los estilos de vida y las enfermedades (129). Este tipo de estudios se extendieron posteriormente a patrones de metilación asociados a enfermedades, comparando los epigenomas de gemelos MZ en los que uno presentaba una enfermedad y el otro no (178). Gracias a las observaciones derivadas de estos trabajos, la metilación del ADN se situó como una herramienta de gran interés en la discriminación de gemelos MZ para el campo forense.

Para el desarrollo de esta metodología se comenzó la búsqueda de marcadores que presentasen diferencias en los patrones de metilación entre pares de gemelos, empleando, principalmente, tecnologías de descubrimiento como las comentadas anteriormente de Illumina (27K (179), 450K (180) y EPIC (181)). Con la búsqueda de estos marcadores comienza a gestarse la que quizás sea la principal limitación de esta aplicación, la diferencia a pequeña escala observada en los patrones de metilación entre gemelos. Esto no quiere decir que no existan diferencias, se observan disimilitudes entre gemelos monocigóticos, pero, aunque significativas en los estudios publicados, no son muy superiores al 10% (180,181). Aun con todo, empleando diversas tecnologías como curvas de fusión de alta resolución (182), pirosecuenciación (183), qPCR (180) y qPCR *MethyLight* (184), se intentaron validar los marcadores identificados. Si las limitadas diferencias observadas no eran desventaja suficiente, tratándose de un biomarcador que presenta cierta variabilidad intrínseca, se observó que los marcadores no eran aplicables a todos los pares de gemelos evaluados, identificándose, en algunas ocasiones, diferencias en los patrones de metilación para los marcadores seleccionados solo en el 12% (183) o en el 18% (179) de los pares de gemelos evaluados. Ante las dificultades encontradas en la resolución de estos sucesos, se planteó la posibilidad de realizar un paso de descubrimiento y validación específico para cada par de gemelos MZ interrogado (180,184). Este procedimiento, aunque factible, es difícilmente abarcable a nivel rutinario y sigue manteniendo la limitación anteriormente comentada. Con estos resultados, por el momento, la discriminación de gemelos monocigóticos está muy lejos de ser aplicable a rutina forense. Los objetivos futuros de la aplicación deben centrarse en la identificación de marcadores que presenten unas diferencias más marcadas y que pueden ser aplicables a un amplio porcentaje de pares de gemelos. Es interesante destacar que en algunos estudios se ha observado que las CpGs diferentemente metiladas, recurrentes en pares de gemelos MZ, estaban enriquecidas en lugares de unión a factores de transcripción asociados al sistema inmune, pudiendo convertirse en un posible objetivo para la búsqueda de nuevas posiciones diferenciadoras, siendo éste un sistema que se especializa en la respuesta al ambiente y cuya expresión génica está regulada en gran medida por la metilación del ADN (185). Si bien es cierto que la discriminación de

gemelos MZ mediante el estudio de la metilación del ADN no parece factible por el momento, gracias a los estudios realizados se ha reforzado el uso de este biomarcador en otras aplicaciones. A lo largo de los trabajos publicados no sólo se observaron diferencias epigenéticas entre gemelos monocigóticos, sino que también se observó que incrementaban con el paso del tiempo, siendo más marcadas en gemelos de mayor edad que en los pares más jóvenes (129,182). El trabajo desarrollado por Fraga et al. destacó la gran influencia del ambiente sobre este biomarcador, ya que estas disimilitudes se acrecentaban si los estilos de vida de los gemelos eran diferentes, siendo más marcadas cuanto más tiempo hubiesen vivido separados y observándose diferencias mínimas en gemelos monocigóticos más jóvenes que vivían vidas análogas y pasaban más tiempo juntos. A su vez, en otros trabajos, se identificaron diferencias asociadas al tipo de tejido evaluado (183,184), siendo, en ocasiones, más significativas que las observadas entre pares de gemelos. Por otro lado, también se observó una limitación del análisis de la metilación del ADN que se iría reforzando con los años, se identificaron diferencias en los patrones de metilación asociadas a la metodología de análisis empleada (180,184).

Gracias a trabajos como estos se fueron definiendo o configurando las diferentes aplicaciones e identificando los problemas asociados al análisis de la metilación del ADN. La discriminación de gemelos monocigóticos puede que esté lejos de ser una realidad, pero el estudio de su epigenoma ha arrojado luz y allanado el camino de otras aplicaciones beneficiosas para el ámbito forense. A continuación, se comentarán en mayor profundidad otras prácticas que han evolucionado mejor con el tiempo, convirtiendo a la metilación del ADN en un biomarcador de gran interés a nivel forense.

1.4.3.2 Identificación de tejidos

En genética forense es habitual enfrentarse a muestras procedentes de diversos tejidos y, en muchas ocasiones, a mezclas de éstos. Una parte importante de la casuística es el conocimiento de la evidencia y su composición, información de gran relevancia tanto a nivel judicial, así como a nivel técnico. La identificación del tejido de procedencia de un resto biológico recogido en una escena de un crimen puede aportar contexto sobre el caso, permitiendo definir el tipo de acto delictivo cometido y facilitando su resolución. Por este motivo, el primer paso en rutina forense se centra en la identificación de tejidos, información que condiciona su resolución al definir el procesado óptimo que debe aplicarse a cada muestra. Desde hace más de un siglo los análisis preliminares han estado muy presentes en el ámbito forense, instaurándose las pruebas enzimáticas e inmunológicas basadas en proteínas y la visualización microscópica de espermatozoides como análisis de referencia para la identificación de tejidos (186,187). Pero estas metodologías conllevan ciertas desventajas. Las pruebas basadas en proteínas, cuya estabilidad es menor que la del ADN, son dependientes de la presencia intacta y funcional de la proteína interrogada, viéndose limitadas ante muestras degradadas. Esto provoca que dichas técnicas sean tejido específicas, siendo necesario decidir qué metodología es más conveniente aplicar en cada situación y para cada muestra. Esta decisión es un punto crucial en el flujo de trabajo, ya que dichas metodologías conllevan la destrucción de la muestra o de parte de ella,

pudiendo comprometer el correcto análisis de la evidencia. A fin de sustituir estas metodologías se han propuesto alternativas como los análisis de ARN mensajero (ARNm) (188) o micro ARN (miARN) (189), su uso aún presenta ciertas limitaciones que deben ser abordadas, como la discrepancias observadas entre diferentes estudios o la falta de protocolos de análisis específicos para análisis forenses (190), pero sigue evaluándose como un posible sustituto a las metodologías tradicionales. Otro biomarcador que han generado interés en los últimos años es la metilación del ADN. Al observarse discrepancias en las pautas de metilación entre tejidos en modelos animales, se planteó la importancia que albergaba en la diferenciación celular (191). Esta relación entre la metilación del ADN y la diferenciación celular se abordó rápidamente en humanos (192,193), observándose patrones de metilación diferenciados entre distintos tejidos. Estos trabajos dieron pie a que, una vez más, una vez asentado el conocimiento, el campo forense comenzase a implementarlo.

La irrupción de esta aplicación en el ámbito forense se produce en 2011 (194), presentando ciertas ventajas sobre las metodologías instauradas. El análisis de la metilación del ADN se realiza sobre el ADN extraído, no necesitando consumir parte de la muestra como los ensayos enzimáticos o inmunológicos, y permitiendo la construcción de *multiplex* que permitan identificar varios tejidos de forma simultánea. Pero la situación inicial del campo limitaría en cierta medida la correcta selección de marcadores específicos de tejido, en concreto para ciertos tejidos de interés forense. Es importante tener en cuenta que los patrones de metilación del ADN de un tejido son dependientes de las poblaciones celulares que lo componen, existiendo tejidos que presentan patrones de metilación estables (sangre y semen) y tejidos, con una mayor complejidad celular (saliva, hisopo bucal, fluido vaginal y sangre menstrual), que presentan patrones de metilación variables dependientes de las proporciones celulares que componen cada muestra (195–197). A ese inconveniente debemos añadir las limitadas bases de datos existentes, donde la variedad en los tejidos analizados brillaba por su ausencia, estando ciertos tejidos infrarrepresentados en cuanto al número de muestras analizadas o siendo directamente inexistentes. Reflejando este escenario, la selección de marcadores estaba posiblemente sesgada observándose una mejor diferenciación de sangre y semen (198–200), mientras que para el resto de los tejidos evaluados, saliva, hisopo bucal, fluido vaginal y sangre menstrual, se obtenían rangos de metilación intermedios, lo que provocaba un solapamiento de los tejidos dificultando su clasificación (198–201). Estas superposiciones entre los patrones de metilación obtenidos y los amplios rangos observados, por ejemplo, porcentajes entre el 14,2% y el 68,5%, se presentan en tejidos que comparten cavidad y que por tanto están en contacto uno con el otro (saliva e hisopo bucal, fluido vaginal y sangre menstrual). Un problema similar se observa en tejidos que comparten poblaciones celulares (células de la sangre en saliva, sangre y sangre menstrual) que, aunque no tienen por qué presentar un rango amplio de valores, pueden proporcionar patrones de metilación similares (200), dificultando su discriminación. Esta situación crea la necesidad de una mejor selección de marcadores específicos de tejido que permitan lograr una diferenciación aun cuando es probable que, para algunos tejidos, parte de los patrones de metilación se compartan. Con este objetivo se establecieron ciertas recomendaciones a fin de garantizar la robustez de los modelos desarrollados; i) presentar idealmente, un nivel de

metilación cercano al 100% (hipermetiladas) o al 0% (hipometiladas) en el tejido de interés y el patrón contrario en el resto, ii) que los patrones de metilación no sean dependientes del sexo, aunque más adelante veremos como condiciones como éstas pueden ser empleadas en favor del análisis y iii) que las posiciones seleccionadas no varíen con la edad.

La aplicación sufrió una mejora en sus resultados tras un cambio de paradigma en la selección de los marcadores de interés. En sus inicios, los modelos generados seleccionaban los marcadores a través de revisiones bibliográficas (198–202) pero, con la llegada de los nuevos paneles de Illumina específicos de metilación, la selección de marcadores se tornó más amplia y específica (203–208), efecto que se vio reflejado en los modelos generados. A lo largo de la historia de esta aplicación en relación con el ámbito forense, se han empleado diferentes metodologías para la validación de los marcadores seleccionados, destacando el uso de enzimas de restricción (194,208,209), pirosecuenciación (198,201–203,210), SNaPshot (200,204–206) y la fusión de alta resolución (207), ofreciendo versatilidad tecnológica. Gracias al uso de posiciones que presentan patrones hiper- e hipometilados se logró construir modelos que permiten diferenciar entre sangre venosa, saliva, sangre menstrual, fluido vaginal y semen (204–206). El desarrollo de modelos de clasificación general de tejidos de interés forense no fue el único acercamiento evaluado empleando este tipo de marcadores, otros trabajos centraron sus esfuerzos en una correcta identificación de un vestigio de gran relevancia en casuística forense, el semen. Por tanto, se desarrollaron modelos de clasificación que permitían identificar semen frente a sangre, hisopo bucal, piel, epitelio vaginal, sangre menstrual y orina (202,209). El trabajo presentado por Wasserstrom et al. (209) arroja resultados prometedores al evaluar muestras mezcla de casos reales, replicando los resultados obtenidos mediante métodos tradicionales ampliamente aceptados como la microscopía en el 96% de los casos analizados. La única limitación que destacan hace referencia a la proporción de contribuyentes. Si la presencia del contribuyente minoritario (presumiblemente donante de semen) es demasiado baja, podría no ser detectada con el modelo basado en metilación del ADN, mientras que empleando microscopía la presencia y observación de un único espermatozoide confirma la presencia de semen. Estas conclusiones extraídas del análisis de muestras reales son consistentes con los resultados observados en otros trabajos que evalúan de forma directa distintas proporciones de mezclas, identificando correctamente el tejido de interés siempre y cuando su contribución a la mezcla sea de al menos un 20% (mezcla 1:5) (206,208). La correcta identificación de semen, sobre todo en mezclas con fluido vaginal, es de gran importancia en casuística forense por lo que recientemente se ha presentado un planteamiento muy interesante para la identificación de tejidos en mezclas. A fin de solventar esta problemática se llevó a cabo el primer acercamiento, en esta aplicación, al uso conjunto de marcadores específicos de tejido en posiciones autosómicas y en posiciones del cromosoma Y (208). Los modelos desarrollados con estos marcadores permiten identificar la existencia o ausencia de contribuyente masculino y si el resto biológico del mismo es procedente de semen o no. A su vez, el conjunto de marcadores de cromosoma Y permiten identificar la presencia de semen en mezclas con *ratio* de 1:100, siempre y cuando la concentración del contribuyente masculino no sea inferior a 200 pg. La propuesta de este trabajo crea nuevas posibilidades técnicas que podrían permitir evaluar

con mayor precisión mezclas complejas, configurando un método que, aun necesitando más evaluaciones, podría sustituir a las técnicas rutinarias. Por otro lado, el estudio de este vestigio biológico ha aportado información muy relevante que, de identificarse, permitiría aportar un mayor grado de informatividad al análisis, observándose diferencias en los patrones de metilación de semen de hombres vasectomizados en comparación con semen de hombres no vasectomizados (200). Si pensamos en la composición celular de las muestras, estos resultados cobran mayor sentido, el semen de un individuo vasectomizado no presenta espermatozoides siendo las células mayoritarias los leucocitos y las células epiteliales, algo que se refleja en la publicación presentando dichas muestras patrones de metilación con una mayor similitud a la sangre menstrual y al fluido vaginal. Este estudio, aunque limitado por el número de muestras y los marcadores seleccionados (que no están pensados para identificar esta diferenciación), abre las puertas a futuros trabajos que busquen discriminar entre esos dos tipos de muestras de semen y, lo más importante, que permitan diferenciar semen vasectomizado de fluido vaginal y sangre menstrual.

Todos estos estudios proporcionan una mayor profundidad a la aplicación que estamos abordando en este apartado, pero también para el biomarcador de interés evaluado. Los modelos desarrollados para la discriminación de tejido han logrado resultados que permiten su aplicabilidad en casuística, ya sea sustituyendo a los métodos tradicionales o como métodos confirmatorios suplementarios. Esta idea se ha reforzado al demostrar la robustez y reproducibilidad de uno de los modelos desarrollados mediante un ejercicio colaborativo entre 12 laboratorios de diversas partes del mundo (211), identificándose correctamente el tejido de procedencia del 97,57% de las muestras analizadas. Pero aún quedan a la espera vestigios que, por su complejidad o poca accesibilidad, no han sido completamente diferenciados o aún no han sido correctamente evaluados, destacando la compleja discriminación entre saliva e hisopo bucal. Recientemente se ha publicado un trabajo que se mueve hacia otras fronteras, abordando la identificación de órganos con metilación del ADN mediante el empleo de marcadores específicos de tejido, logrando discriminar entre epidermis, dermis, corazón, musculo esquelético, sangre, riñón, cerebro, pulmón e hígado (212). Uno puede encontrar símiles entre la investigación y las aventuras, recordando las bonitas palabras de un viejo *hobbit* de Bolsón Cerrado, Bilbo Bolsón *"¿Las aventuras nunca tienen un final? Supongo que no. Alguien más siempre tiene que continuar la historia."*. Y es que esta historia, la historia de la metilación del ADN, continúa con esta y otras aplicaciones, pero todas ellas deben tener presente la comentada en este punto. El tejido cobra una importancia capital en cualquier aplicación asociada a este biomarcador, focalizándose, a un nivel más concreto, en las poblaciones celulares que componen la muestra.

1.4.3.3 Estimación de la edad

La edad es un constructo social, un valor numérico que empleamos para medir un concepto tremendamente complejo, el envejecimiento. Por tanto, para poder hablar de estimación de la edad, tanto en el ámbito clínico como forense, debemos comprender de forma somera los mecanismos subyacentes, para así poder definir la mejor estrategia de estudio. El

envejecimiento implica un deterioro fisiológico de nuestro cuerpo, conformando el principal riesgo para la mayor parte de patologías humanas (213). El estudio de esta condición se ha planteado desde diversos puntos de vista, desde la evaluación del genoma de personas centenarias identificando genes asociados a la longevidad, hasta estudios en gemelos monocigóticos evaluando modificaciones epigenéticas, alteraciones que podrían ocurrir con mayor frecuencia debido a que sus consecuencias para la supervivencia son menos drásticas (214). El envejecimiento ha demostrado ser un proceso muy heterogéneo, describiéndose como un mosaico complejo resultado de la interacción de diversos factores genéticos, epigenéticos, ambientales y estocásticos (214). Esta complejidad hace que sea muy complicado definir un fenotipo asociado a este proceso, presentando cada individuo patrones y velocidades de envejecimiento diferentes, siendo incluso discrepante entre los distintos órganos de un mismo individuo (215). Con los primeros estudios se plantearon dos teorías biológicas para explicar el envejecimiento: las estocásticas y las de desarrollo genético (215). La primera de ellas hacía referencia al daño aleatorio que sufren las moléculas con un papel crítico en rutas metabólicas, comprendiendo: acumulación de mutaciones somáticas, procesos postraduccionales (oxidación y glicosilación) que producen alteraciones en las proteínas y acumulación de radicales libres que derivan en estrés oxidativo. Por otro lado, las teorías de desarrollo genético se basan en los procesos de senescencia y degeneración paulatina del sistema inmunitario. Ninguna de las dos teorías planteadas fue descartada y la realidad es que ambas configuraban parte de los mecanismos implicados en el proceso de envejecimiento, hecho resaltado con la publicación de los rasgos distintivos del envejecimiento (*Hallmarks of aging*) (213). Este trabajo tuvo un gran impacto, definiendo y agrupando factores asociados al envejecimiento que proporcionaba una guía con los diferentes puntos de mayor repercusión del campo. Teniendo esto en cuenta, es conveniente comentar, aunque sea de forma general, las alteraciones en nuestro organismo que fomentan el proceso de envejecimiento y que engloban estos nueve rasgos distintivos. A nivel genético está presente la inestabilidad genómica tanto nuclear como mitocondrial, observándose una asociación entre las mutaciones somáticas y el envejecimiento (216,217). Entre ellos, las alteraciones observadas en el ADN mitocondrial son las más interesantes por la limitada eficiencia de los mecanismos de reparación en comparación con el ADN nuclear (218). La disfunción de estos orgánulos se traduce en la pérdida de eficiencia de la cadena respiratoria, provocada por una reducción en la generación de adenosín trifosfato (ATP) (219). Este proceso presenta efectos asociados al envejecimiento tales como deterioro de las mitocondrias, daño celular y apoptosis. Asociadas a mutaciones, las producidas en los genes de la hormona de crecimiento y del factor de crecimiento insulínico engloban, por sí mismas, otro rasgo denominado desregulación de la detección de nutrientes, observándose una relación entre dieta, longevidad y envejecimiento (220). A parte de mutaciones, la inestabilidad genómica hace referencia a otro de los rasgos distintivo de este proceso, uno muy conocido, el acortamiento de los telómeros (221). Las telomerasas son las encargadas de completar las secciones terminales de las cadenas lineales de ADN, pero las células somáticas de mamíferos no expresan esta enzima, produciéndose una progresiva y acumulativa pérdida de los telómeros. Si nos quedamos cerca del ADN nuclear o, más bien, por encima, nos encontramos con otro de los rasgos propuestos, las alteraciones epigenéticas (222). Un punto de gran interés a nivel clínico

de estas marcas es su reversibilidad, lo que podría abrir las puertas al diseño de tratamientos antienvejecimiento (223). Por otro lado, otro de los rasgos definidos está asociado con el correcto plegamiento de las proteínas, la proteostasis (224). Las proteínas que no se pliegan correctamente deben ser restauradas o eliminadas, cuando esto no ocurre se acumulan en el organismo observándose una asociación entre su acumulación y el envejecimiento. Por último, los tres rasgos restantes están asociados al nivel celular, siendo la senescencia, la extenuación de las células madre y la comunicación intercelular alterada. El primero de ellos es un proceso beneficioso que ocurre en nuestro organismo, previene la propagación de las células dañadas deteniendo su división celular y siendo eliminadas por el sistema inmune. El problema surge cuando existe un fallo en el sistema de eliminación y renovación que recupera el número de células perdidas, relacionándose estos errores en organismos envejecidos a la perdida de eficiencia de los sistemas con la edad. Otro de los rasgos está relacionado con esta pérdida de funcionamiento, la extenuación de las células madre. El decaimiento de la capacidad regenerativa de los tejidos es una de las principales características del envejecimiento y el rejuvenecimiento de células madre se ha planteado como un interesante método para combatir este proceso (225). El último de los rasgos distintivos del efecto de la edad es la comunicación intercelular alterada, produciendo las células señales incorrectas como fenotipos proinflamatorios (226). Estos procesos pueden ser producto de alguna de los rasgos citados anteriormente como, por ejemplo, disfuncionalidad del sistema inmune o secreción de citoquinas proinflamatorias por las células senescentes.

Todo este conocimiento configuró el campo del envejecimiento y su estudio orbitó alrededor de estas alteraciones. Con los años se aportaron un mayor número de pruebas que asociaban los rasgos descritos al efecto en nuestro organismo del paso del tiempo o se identificaron nuevos rasgos característicos de este proceso. Esto puede verse reflejado en la revisión presentada en 2023 por López-Otín et al. (227) donde se profundiza en el conocimiento generado para los nueve rasgos distintivos presentados en 2013 y se añaden tres nuevos que se han identificado durante los últimos diez años; desactivación de la macroautofagia (antes incluida dentro de la proteostasis), inflamación crónica (antes parte de la comunicación intercelular alterada) y la disbacteriosis (alteración del microbioma intestinal). Gracias a estos trabajos se ha configurado un ambiente muy propicio para estudiar directamente estas alteraciones con el objetivo de predecir la edad biológica o cronológica de un individuo. Estos no serían los primeros planteamientos o metodologías empleadas en la predicción de dicho rasgo, pero proporcionaría un nuevo acercamiento que cambiaría el campo de la predicción de la edad, siendo beneficioso tanto a nivel clínico como forense.

1.4.3.4 Primeras aproximaciones en la estimación de la edad

La estimación de la edad es un campo que a nivel clínico y forense presenta una diferencia conceptual muy grande que hace que los modelos no sean intercambiables. En el ámbito clínico, el objetivo busca predecir la edad biológica estudiando las alteraciones que afectan al envejecimiento del organismo de un individuo, con el fin último de recomendarle medidas o tratamientos que puedan revertir, de ser necesario, el proceso. Esto es una consideración a nivel

teórico y, si bien es cierto que cada vez se estudian más tratamientos con el objetivo de revertir el envejecimiento, su aplicación rutinaria aún está lejos de ser una realidad asumible. Por otro lado, a nivel forense, el interés reside en la edad cronológica, la cual ofrece un dato que, junto con otros, confiere una capacidad individualizante a la muestra. La predicción de la edad busca aportar información diferenciadora que permita reducir el número de sospechosos en casos criminales, asistir en la identificación de restos humanos en desastres de masas, en el estudio de restos arqueológicos o mejorar la predicción de características físicas asociadas al envejecimiento, como el encanecimiento o la calvicie masculina.

La estimación de la edad ha sido abordada empleando diversas metodologías, estudiándose inicialmente fases fisiológicas que presentaban características comunes para todas las personas, como por ejemplo el desarrollo físico, esquelético o dental. En estos estudios se definían estados de estas características asociados a sexo y edad, conformándose conjuntos de referencia que se empleaban para definir la edad de un sujeto concreto. En los exámenes óseos y dentarios se definían tres fases de estudio que presentaban distintas características evaluables mediante estudios radiográficos: crecimiento y desarrollo, equilibrio y senescencia (228). Mientras que para los huesos se estudiaba el tamaño, la forma o el estado de osificación de las placas epifisarias (229,230), en los estudios dentales se evaluaba la erupción y mineralización de las piezas mediante inspección visual y ortopantomografía (231), respectivamente. Esta metodología para la predicción de la edad no presentaba una precisión óptima y en algunos casos requería de la exposición a rayos-X por lo que no tardaron en explorarse métodos alternativos.

Como hemos visto en el apartado anterior, el envejecimiento está asociado a modificaciones graduales que se producen en las biomoléculas de nuestro cuerpo, por tanto, el estudio molecular de estas alteraciones se ha evaluado como posible mecanismo de estimación de la edad, agrupándose en 5 grupos: i) deleciones de ADN mitocondrial, ii) acortamiento de telómeros, iii) productos finales de glicación avanzada (AGEs), iv) racemización del ácido aspártico y v) *signal-joint T-cell receptor excision circles* (sjTRECs). Comentemos brevemente estos procesos, que quizás nos suenen de los rasgos distintivos del envejecimiento, y sus limitaciones a la hora de aplicarlos en el ámbito forense. El primero de ellos se basa en el daño progresivo, producido por radicales libres, que sufre el ADN mitocondrial (ADNmt) a lo largo de la vida, por tanto, se estudia la edad en relación con la acumulación de deleciones (232). Su aplicación forense se ve condicionado por limitaciones metodológicas y por una escasa capacidad predictiva (233). El segundo caso, es el acortamiento telomérico (234), pero la necesidad de una gran cantidad de ADN y su poca reproducibilidad, el acortamiento es dependiente del individuo; hacen que su implementación forense sea compleja (233). El tercer grupo centra su atención en los cambios de color producidos en los tejidos por la acumulación de proteínas procedentes de la glicación de proteínas y su relación con el envejecimiento (235). En este caso, la heterogeneidad de este grupo de compuestos y la falta de metodologías estandarizadas ha limitado enormemente su evaluación en el ámbito forense (233). El siguiente grupo de técnicas centra su estudio en la racemización, proceso químico que convierte enantiómeros puros (L o D) en una mezcla de ambos, y presenta, para el ácido aspártico, los

mejores resultados de predicción de las cinco categorías (error de ± 3 años) (236). En mamíferos durante la síntesis de proteínas se incorporan exclusivamente L-aminoácidos, produciéndose con la edad la racemización y aumento de los niveles de D-aminoácidos. Aunque la técnica ha demostrado una buena predicción, las limitaciones presentadas han hecho inviable su aplicación forense ya que es una técnica destructiva y está limitada a un único tipo de muestra, la dentina (233). El último de los grupos evalúa las moléculas de ADN circulante producto del reordenamiento somático que sufren los linfocitos-T inmaduros (237). Esta técnica molecular es la más viable pero sus errores de predicción son elevados (± 9 años) y su uso está restringido a muestras de sangre. Ante este escenario las expectativas en esta aplicación forense flaqueaban, pero el descubrimiento de marcadores epigenéticos correlacionados con la edad abrió un nuevo camino que con el tiempo fue explorado y aprovechado.

La asociación entre la metilación del ADN y la edad fue publicada en los años 80-90, identificándose una asociación global entre envejecimiento y patrones de metilación (238,239). El paisaje epigenético vinculado a este biomarcador con respecto a la edad sigue una tendencia general de hipometilación con el paso del tiempo, partiendo los patrones de metilación al principio de nuestras vidas desde un nivel elevado (hipermetilación) (240). Es preciso recalcar que esta observación es general, identificándose otras tendencias en puntos específicos del genoma, en concreto ciertos promotores asociados a islas CpGs presentan hipermetilación con la edad (241,242). Para entender estos patrones es preciso evaluar cómo pueden afectar al organismo. Mientras que los patrones de hipometilación no parecían seguir ninguna pauta concreta, la hipermetilación se concentraba en posiciones asociadas a procesos de expresión génica, metabolismo de nucleótidos y morfogénesis (243). Estas predisposiciones tenían una repercusión mayor y rápidamente la relación entre metilación del ADN y edad comenzó a coger forma, surgiendo los conceptos edad biológica y edad cronológica asociados a este biomarcador. Al ser la hipometilación un proceso pasivo mediado, principalmente, por efectos estocásticos y ambientales se vinculó con la edad biológica, por otro lado, la hipermetilación al asociarse con un proceso activo programado de envejecimiento se relacionó con la edad cronológica (243). Como resultado de estas afirmaciones se definieron dos nuevos conceptos asociados a las modificaciones que produce en el epigenoma el envejecimiento: deriva epigenética y reloj epigenético. La deriva epigenética se produce por la acumulación estocástica de cambios producidos durante los procesos de transmisión y mantenimiento de la metilación a lo largo del genoma (244). Este proceso fue observado gracias a estudios realizados con gemelos monocigóticos, algo que comentamos anteriormente, para los que se observaban patrones de metilación casi idénticos al inicio de sus vidas y, por el contrario, a edades ancianas las diferencias observadas eran mucho más pronunciadas (129). Por otro lado, el reloj epigenético hace referencia al otro tipo de modificaciones que se observan, aquellos cambios producidos con la edad en lugares específicos del genoma, siendo estos cambios progresivos y comunes entre individuos (245). Gracias a todos estos estudios se configuró un ambiente propicio que permitió a la comunidad científica moverse hacia una época en la que convivían estudios de descubrimiento y el desarrollo de modelos de predicción de la edad. Los patrones

de metilación asociados a la edad comenzaban a conocerse y la correcta aplicación de ese conocimiento sería clave para predecir la edad de un individuo.

1.4.3.5 Modelos epigenéticos de predicción de la edad

a) Conceptos generales

La construcción de modelos de predicción es siempre una tarea desafiante, la correcta evaluación de los datos disponibles y del acercamiento estadístico empleado son piezas fundamentales en su desarrollo, pero en ocasiones se deben tener en cuenta otros factores. El proceso de modelado en cualquier aplicación asociada a la metilación del ADN presenta unas características específicas ineludibles que, siempre que sea posible, deben ser investigadas e informadas a la hora de presentar el modelo. Uno de estos factores está intrínsecamente relacionado con la variable estudiada, la edad. En este caso vamos a centrar nuestra atención no solo en la edad, sino más bien en el rango de edad analizado, ya que puede definir el valor del modelo y los resultados obtenidos. Para desarrollar un buen modelo es preciso estudiar un amplio rango de edades, siendo dicho rango el que define las edades predecibles por el modelo, por tanto, para emplearlo correctamente, aquellos individuos con edades fueran del rango estudiado no deberían ser analizados con el modelo. ¿Cómo identificamos esto en una muestra desconocida? Evaluando si los valores de metilación de los marcadores estudiados se encuentran dentro del rango obtenido para las muestras analizadas en el conjunto de entrenamiento. Pero hay que tener en mente una tendencia observada debido a la naturaleza del biomarcador analizado que puede afectar en gran mediad a los resultados obtenidos, la presencia de mayores diferencias en los niveles de metilación en personas ancianas de la misma edad (129). Al acumularse alteraciones estocásticas y ambientales a lo largo de la vida, se ha observado como individuos de la misma edad presentan patrones de metilación diferentes, esto, en las representaciones gráficas de los modelos de edad, se presenta como un cono de dispersión. Los individuos más jóvenes presentan patrones de metilación más agrupados, observándose una dispersión creciente a medida que aumenta la edad, siendo la dispersión más pronunciada en edades avanzadas. Por tanto, la presencia de un mayor número de individuos ancianos en el conjunto de entrenamiento o de validación de un modelo generará un mayor error de predicción. Esta situación se ha afrontado equilibrando la representación de todas las edades en los modelos, con el objetivo de representar a toda la población y no crear modelos sesgados sin individuos ancianos. Otro de los factores de importancia vital para cualquier estudio asociado con la metilación del ADN, como ya hemos visto en el apartado anterior, es el tejido analizado. Al ser este biomarcador especifico de tejido, la selección de marcadores o la generación de modelos debe realizarse en base a cada tejido de interés. Esto no imposibilita la generación de modelos multitejido, pero su construcción se torna más compleja y demandante. Ante esta posibilidad es importante destacar, sobre todo en el ámbito forense, que estos modelos multitejido no son aplicables a muestras que presenten mezcla de tejidos. Ésta se presenta como una limitación de gran repercusión a nivel forense donde es habitual enfrentarse a este tipo de muestras, siendo necesario modelos de identificación de tejido que permitan discriminar si una muestra presenta o no una mezcla para asegurar la correcta

aplicación de un modelo basado en metilación del ADN. Esta restricción está asociada a la medición del biomarcador, donde los valores obtenidos presentan simplemente el nivel de metilación observado en una muestra en uno o varios marcadores específicos, por tanto, las mezclas presentarán esta proporción, pero no nos permitirán discernir el grado de contribución de cada tejido. Dicha barrera no es exclusiva de mezclas de tejidos y la metilación del ADN de muestras con más de un contribuyente sufre la misma suerte, no permitiendo identificar la aportación de cada individuo a la mezcla. La solución a esta limitación aún debe ser abordada, pero existen otros factores que deben tenerse en cuenta durante el proceso de desarrollo.

La tecnología de análisis de metilación empleada para la generación del modelo ha cobrado importancia en los últimos años, observándose diferencias en los valores de metilación obtenidos para los mismos individuos en las mismas posiciones CpG empleando distintas plataformas (211,246–248). Este hecho crea la necesidad de analizar las nuevas muestras, muestras dubitadas para las que se quiere predecir la edad, en la plataforma empleada durante la construcción del modelo utilizado. Para abordar este problema se ha planteado la construcción de modelos de transformación, o introducción, en los modelos de predicción de edad, de una variable que represente la plataforma empleada o el cálculo de factores de corrección entre plataformas extendiendo la aplicabilidad de los modelos ya desarrollados (246–248). La comparación directa de las diferentes tecnologías con mayor repercusión en el campo ha sido abordada empleando un mismo conjunto de muestras (249). Las técnicas cuantitativas evaluadas en este trabajo, EpiTYPER, pirosecuenciación y MiSeq®, presentaron altos niveles de similitud en los patrones de metilación obtenidos, haciendo factible la intercambiabilidad de los datos tras un proceso de validación específico. Por otro lado, SNaPshot, una tecnología semicuantitativa, presentó diferencias más marcadas posiblemente derivadas de las diferentes señales obtenidas por los fluorocromos empleados para la detección de los patrones de metilación. Es importante destacar que las diferencias parecen estar asociadas al marcador, lo que conlleva que estos resultados no puedan ser extrapolados, siendo necesario evaluar o transformar los valores de una plataforma en otra teniendo en cuenta de forma individual cada uno de los marcadores presentes en el modelo analizado. En los últimos años, la metodología de SNaPshot ha sido evaluada de forma individual entre las plataformas empleadas para su análisis (250). Teniendo en cuenta las diferencias asociadas al equipo empleado se ha concluido que el acercamiento óptimo conlleva generar un modelo de predicción específico para cada secuenciador, pero, por suerte, existen otros planteamientos que pueden ser aplicados para mantener vivos los modelos desarrollados. La realidad es que el reanálisis y reconstrucción de los modelos de predicción para cada nuevo secuenciador es un proceso largo y costoso que no todos los laboratorios pueden permitirse. El uso de ecuaciones de ajuste ha demostrado ser una solución factible que proporciona buenos resultados y permite extender el uso de los modelos a nuevas plataformas. Si bien es cierto que aún no existe un modelo de predicción de la edad definitivo, es importante tener en cuenta esta variable para asegurar la longevidad de los modelos desarrollados, sobreviviendo a la obsolescencia de las tecnologías que los han generado.

Estos modelos existen no solo gracias a las técnicas de análisis empleadas, sino también gracias a la metodología estadística utilizada. La decisión de qué modelo estadístico emplear tiene una repercusión importante y debe plantearse con cuidado antes de la generación de cada modelo. A lo largo de los años se han empleado diversas técnicas entre las que podemos encontrar modelos de regresión lineal, modelos de regresión cuantil y técnicas de aprendizaje automático como redes neuronales, bosques aleatorios y vectores de soporte. La elección del tipo de modelo empleado es de gran importancia porque puede tener repercusiones a la hora de aplicarlo a nuevos conjuntos de datos, o lo que es lo mismo, muestras dubitadas. Una tendencia general en el campo es el uso de modelos de regresión lineal para la generación de modelos de predicción de la edad, algo que quizás no sea del todo recomendado. Los modelos de regresión lineal tienen ciertas hipótesis que o se comprueban o se asumen al generar el modelo, entre ellas encontramos: distribución normal, homocedasticidad de los datos (varianza constante), independencia estadística entre variables y linealidad (251). Algunas de estas asunciones tienen mayor repercusión que otras, pero la falta de homocedasticidad, o presencia de heterocedasticidad en los datos, puede repercutir en una pérdida de precisión a la hora de predecir muestras externas al conjunto de datos de entrenamiento. Por desgracia, la metilación del ADN en marcadores de la edad presenta, de forma general, una distribución heterocedástica ya que, como comentamos anteriormente, se observan diferencias menores entre individuos jóvenes y diferencias mayores entre individuos ancianos de la misma edad. Si bien es cierto que esto no impide el uso de esta metodología estadística, a la vista están los buenos resultados obtenidos con modelos lineales, es importante tener esto en mente y evaluar los datos en consecuencia. Esta recomendación se extiende a todos los modelos estadísticos ya que todos tienen sus ventajas e inconvenientes dependiendo del conjunto de datos y tipo de variables analizadas. Por ejemplo, entender las máquinas de aprendizaje que tan de moda están en estos momentos es muy complejo y si bien es cierto que pueden proporcionar muy buenos resultados, pueden generar sobreestimación si el conjunto de entrenamiento no es lo suficientemente grande.

Durante la selección de marcadores y la recogida de muestras existen otros elementos que pueden repercutir en la correcta predicción de la edad. Como ya hemos comentado y veremos en más detalle en apartados posteriores, la metilación del ADN se ve influenciada por factores ambientales, entre los que podemos destacar los estilos de vida y las enfermedades. Estas variables no sólo tienen un efecto global o específico en el epigenoma, sino que se ha visto que están asociados a un aceleramiento o desaceleramiento del envejecimiento epigenético (252–256). Su tratamiento durante el desarrollo de un modelo de predicción varía según el objetivo de este, observándose perspectivas diferentes al hablar de modelos clínicos y forenses. Aquellos con carácter clínico integran estilos de vida y enfermedades para plasmar los efectos de aceleración de dichas variables, permitiendo predecir la edad biológica del paciente (257–259). Este planteamiento es completamente contrario a lo observado en el ámbito forense, donde se busca predecir la edad cronológica seleccionándose marcadores correlacionados con la edad que no presenten alteraciones derivadas de factores ambientales o enfermedades. Con esto en mente se propuso una división entre modelos diseñados para predecir la edad cronológica,

llamados relojes epigenéticos cronológicos o relojes de edad forenses, y la biológica, llamados relojes epigenéticos biológicos (260). Por tanto, la evaluación del efecto de los estilos de vida y las enfermedades en los marcadores seleccionados para los modelos forense ha sido abordada en algunos casos (261,262). La disponibilidad de esta información puede ser limitante a la hora de evaluar estos factores, por lo que es habitual que no puedan ser tenidos en cuenta en muchos estudios. Aun así, siempre que sea posible, es recomendable comprobar si las posiciones seleccionadas para la construcción de un modelo pueden verse afectadas, ya que podrían tener un efecto en las predicciones.

La aplicación de la predicción de la edad en el ámbito forense despegó tras la publicación del modelo multitejido desarrollado por Horvath en 2013 empleando 353 CpGs (263). Si bien es cierto que marcó un antes y un después en el campo, su aplicabilidad forense estaba lejos de ser una realidad y se apostó por la construcción de modelos independientes para cada tejido empleando el menor número de marcadores posibles. Este modelo se considera la piedra *rosetta* y posiblemente uno de los trabajos que más han impulsado un campo que hoy en día sigue en auge. A nivel forense, en los últimos años, se han abordado diversos tejidos de interés y se han generado modelos empleando diferentes metodologías, marcadores y evaluaciones estadísticas reduciendo sustancialmente los errores de predicción obtenidos. En el marco de todos estos avances, el desarrollo de estudios inter-laboratorio ha demostrado el interés de la comunidad por la aplicabilidad de los modelos desarrollados en laboratorios de todo el mundo (211,264,265). Comentemos los modelos desarrollados teniendo en cuenta que, en ocasiones, los parámetros estadísticos aportados en los diferentes artículos publicados no son directamente comparables.

b) Modelos de predicción de la edad en sangre

La sangre es uno de los tejidos más estudiados en genética forense. Antes de comenzar a evaluar los avances en los modelos desarrollados en base a muestras de sangre, debemos detenernos brevemente para analizar el tejido de interés. Anteriormente se ha insistido en la importancia del tejido a la hora de trabajar con metilación del ADN, pero hay tejidos que no podemos considerar un conjunto homogéneo porque están compuestos por diferentes poblaciones celulares. La sangre total es uno de estos casos, su composición y proporción heterogéneas, así como la obtención de perfiles de metilación específicos en células mononucleares de sangre periférica y granulocitos (266), pueden introducir variabilidad en los resultados observados. Sin embargo, se ha identificado que los patrones de metilación de posiciones CpG correlacionadas con la edad en sangre no presentan una variabilidad significativa entre células sanguíneas de individuos sanos (263), por tanto, los modelos de predicción de edad forenses desarrollados toman la sangre total como un tejido homogéneo. Una vez comprendemos el tejido de estudio, es recomendable entender el ambiente de publicaciones que se ha generado con los años, sobre todo con el objetivo de comparar y valorar los modelos desarrollados. Durante el desarrollo de esta aplicación se han evaluado diversos acercamientos estadísticos generándose con ellos múltiples parámetros de análisis: R^2, error absoluto medio (MAE_{media}), error absoluto mediano ($MAE_{mediana}$), error cuadrático medio (RMSE) y el porcentaje de clasificaciones correctas. Ante este paisaje tan diverso, el problema más destacable es la discordancia en los estadísticos aportados en las publicaciones para evaluar los modelos desarrollados. Aun así, existe un parámetro estadístico que siempre está presente y cuyo uso se ha generalizado, configurándose como valor de referencia en la evaluación de los modelos de predicción de la edad en el ámbito forense, el MAE. Por tanto, centraremos nuestra atención en este parámetro a la hora de evaluar los modelos de edad, siendo esta decisión extensible al resto de tejidos evaluados en otros apartados. Una rápida mirada al horizonte nos presenta una amplia multitud de modelos de predicción de la edad basados en muestras de sangre. Este hecho complica la integración de todos ellos en el texto, por tanto, en la Tabla 1, presentada a continuación, se recoge la información más relevante de los principales modelos desarrollados en base a muestras de sangre en el ámbito forense.

Tabla 1. Resumen de los principales modelos de predicción de la edad en base a muestras de sangre. Se aporta la referencia de la publicación del modelo, año de publicación, genes en los que se encuentran los marcadores empleados, metodología técnica utilizada y error obtenido. El número de genes no representa el número de marcadores empleado en la construcción del modelo.

Referencia	Año publicación	Genes empleados	Metodología	Error (MAE)
Weidner et al. (267)	2014	*ITGA2B* *ASPA* *PDE4C*	Pirosecuenciación	± 5,4 años
Zbiec-piekarska et al. (268)	2015	*ELOVL2* *C1orf132* *TRIM59* *KLF14* *FHL2*	Pirosecuenciación	± 4,5 años
Zbiec-piekarska et al. (269)	2015	*ELOVL2*	Pirosecuenciación	± 5,03 años
Park et al. (270)	2016	*ELOVL2* *ZNF423* *CCDC102B*	Pirosecuenciación	± 3,16 años
Freire-Aradas et al. (271)	2016	*ELOVL2* *ASPA* *PDE4C* *FHL2* *CCDC102B* *C1orf132* cg07082267 (sin gen)	EpiTYPER	± 3,07 años
Thong et al. (272)	2017	*ELOVL2* *TRIM59* *KLF14*	Pirosecuenciación	± 3,3 años
Vidaki et al. (273)	2017	*CSNK1D* *C21orf63* *CASC4* *SSRP1* *FXN* *P2RXL1* *RASSF5* *ERG* *TRIP10* *FZD9* *KLF14* *NR2F2* *VGF* *NHLRC1* *SCGN* *C19orf30*	MiSeq®	± 3,8 años
Naue et al. (274)	2017	*ELOVL2* *DDO* *F5* *GRM2* *HOXC4* *KLF14* *LDB2* *NKIRAS2* *RPA2* *SAMD10* *TRIM59* *MEIS1-AS3* *ZYG11A*	MiSeq®	± 3,2 años
Jung et al. (275)	2019	*ELOVL2* *FHL2* *KLF14* *C1orf132* *TRIM59*	SNaPshot	± 3,5 años
Wozniack et al. (276)	2021	*ELOVL2* *MIR29B2CHG* *KLF14* *FHL2* *TRIM59* *PDE4C*	MiSeq®	± 3,2 años
Aliferi et al. (262)	2022	*ELOVL2* *FHL2* *EDARADD* *ARHGAP22* *RASSF5* *CNTNAP2* *MIR29B2CHG* *CDC102B* *TRIM59* *LDB2*	MiSeq®	± 3,6 años
Han et al. (277)	2022	*ELOVL2* *FHL2* *C1orf132* *CCDC102B* *KLF14* *SYNE2* *TRIM59* cg26947034 (sin gen)	SNaPshot	± 2,88 años

El primer modelo forense en base a este tejido se desarrolló en 2014 (267), presentando un error medio absoluto (MAE_{media}) de ± 5,4 años. Este primer error no fue excesivamente elevado, pero a partir de este punto los errores de los modelos, como se puede observar en la Tabla 1, bajarían rápidamente presentándose dos años después un modelo cuyo error era cercano a los ± 3 años (270). Los modelos publicados hasta la fecha no reducirían los errores más allá de cifras cercanas a ésta. Este límite no debería percibirse como algo negativo, el gran trabajo de la comunidad científica identificó diversos marcadores con un alto nivel de correlación que proporcionaron rápidamente resultados muy informativos, de ser aplicados a casos forenses. A su vez, la predicción de la edad en base a muestras de sangre presenta una gran robustez al mostrar resultados consistentes empleando diferentes marcadores, técnicas y metodologías estadísticas. Gracias a todos estos trabajos se ha enriquecido enormemente el campo y no se ha observado un acercamiento metodológico predominante que pueda limitar la aplicación de modelos de predicción de la edad en laboratorios de todo el mundo. Entre las técnicas empleadas, a lo largo de estos últimos diez años de investigación, encontramos principalmente las descritas anteriormente, pirosecuenciación (267–270,272), EpiTYPER (271), SNaPshot (275,277) y MPS (262,273,274,276). Es interesante destacar el cambio de tendencia que se ha observado con el paso de los años. En los inicios, la gran mayoría los modelos se desarrollaron empleando pirosecuenciación, pero, con el tiempo, las ventajas de tecnologías como SNaPshot y MPS han ido ganando importancia en el ámbito forense relegando a la precursora a un puesto más secundario. En relación con los marcadores, como se observa en la tabla, hay un gen predominante que prácticamente está presente en la totalidad de los modelos, *ELOVL2*, gen asociado con la actividad de elongación de los ácidos grasos. Un estudio funcional de este marcador en modelos animales ha demostrado una correlación entre este *locus* y el envejecimiento en retina (278). Esta relación se extendió al estudio de la metilación del ADN, observándose que patrones de envejecimiento en la retina de ratones estaban asociados con el silenciamiento y, por tanto, hipermetilación de este gen (279).

Durante todo este proceso, los modelos generados se han empleado para evaluar otras aproximaciones de interés forense, relacionadas con su aplicabilidad, robustez o sensibilidad. Las situaciones forenses suelen presentar muchas incógnitas relacionadas con el vestigio analizado. Asociado a la predicción de la edad, el parámetro que queremos evaluar puede encontrarse en un amplio rango. Muchos de los modelos desarrollados, presentes en la Tabla 1, han evaluado amplios rangos de edades, analizando tanto muestras de menores como de mayores de edad (262,268–270,272,273,277,280). Pero el grueso de las muestras analizadas pertenece a individuos mayores de edad, estando los menores poco representados. El escaso número de individuos menores de edad y la falta de representación de ciertas edades en ese rango pueden condicionar en gran medida las predicciones obtenidas para muestras de individuos de esas edades. Por tanto, el estudio de los patrones de metilación de la edad de menores cobra gran importancia para conocer su tendencia y proporcionar una mayor robustez a los modelos generados. La construcción de modelos de predicción de la edad en menores ha sido planteada (281), identificándose marcadores correlacionados con este rango de edad y construyéndose un modelo con un error de ± 0,94 años. Este bajo error, en comparación con el

resto de los modelos, está asociado a la baja variabilidad inter-individual que presentan los patrones de metilación en edades tempranas, en contraposición con lo que se observa en ancianos. Pero quizás lo más destacable sea la identificación de tendencias diferentes entre menores y mayores de edad, observándose una mayor correlación con la edad en menores y una estabilización de los patrones de metilación a partir de la adolescencia. Este descubrimiento destaca la importancia del empleo de un conjunto de muestras que abarque el mayor rango de edad posible con una correcta representación de todas las edades, ya que podría ser vital a la hora de enfrentarnos a una muestra desconocida.

Por otro lado, la predicción de la edad en sangre se llevó más allá del uso de sangre fresca, generándose modelos con sangre de fallecidos (282,283) o evaluando la aplicabilidad de modelos basados en sangre fresca en muestras *post mortem* (284). Los modelos generados con muestras de fallecidos o combinando muestras de éstos con muestras de sangre fresca presentaron resultados bastante diferentes (Correia Dias et al. ± 6,08 años y Bekaert et al. ± 3,75 años, respectivamente), lo que no permite obtener una conclusión clara dejando la puerta abierta a nuevos estudios. Por otro lado, al predecir las edades de un conjunto de muestras de fallecidos empleando un modelo construido con muestras de sangre fresca (MAE: ± 5,32 años) los errores obtenidos fueron de ± 9,72 años (284), presentando una posible discordancia destacable entre los patrones de metilación de los tipos de muestra analizados.

Este acercamiento alternativo no fue el único evaluado. Teniendo en cuenta la rutina forense, se analizaron otro tipo de muestras muy comunes en casos criminales, las manchas de sangre. Analizando sangre fresca y manchas de sangre de los mismos individuos no se observaron diferencias significativas en las edades predichas, edades que se mantuvieron concordantes tras el almacenamiento de las manchas de sangre hasta cuatro meses a temperatura ambiente (280). A su vez, a fin de evaluar en mayor profundidad la estabilidad de la metilación del ADN de este tipo de muestra, se analizaron manchas de sangre almacenadas hasta un máximo de 15 años, prediciéndose correctamente la edad independientemente del tiempo de almacenamiento (269). Por otro lado, teniendo en cuenta la limitada cantidad de muestra disponible en muchas ocasiones en el ámbito forense, se realizaron estudios de sensibilidad con el objetivo de definir la cantidad mínima de ADN genómico de partida que proporciona patrones de metilación estables. Los primero estudios que evaluaron la sensibilidad observaron que para cantidades inferiores a diez nanogramos se obtenían mayores desviaciones entre los valores, sugiriéndose el uso de un mínimo de 10 ng (285) o 20 ng (276) de ADN genómico de partida. En los últimos años se ha demostrado que los patrones de metilación pueden mantenerse estables con un mínimo de cinco nanogramos de ADN genómico, manteniéndose, por tanto, los errores de predicción obtenidos (262), acercando cada vez más la sensibilidad de los modelos desarrollados a la realidad forense.

Por último, teniendo en cuenta que en muchas ocasiones no podemos estar seguros de a qué tejido pertenece la muestra recibida, se ha evaluado la aplicabilidad de los modelos generados a otros tejidos de interés forense como saliva (273,285), diente (282), cartílago y músculo (276). Es importante tener en cuenta que la mayoría de dichas evaluaciones son

preliminares, analizan un bajo número de muestras, o presentan errores muy elevados como para ser aplicables (entre ± 7,3 y ± 17,1 años). Solamente el modelo de sangre desarrollado por Vidaki et al. (273) presenta un error similar para sangre y para las muestras de saliva empleadas como conjunto de validación (± 3,3 años y ± 3,2 años, respectivamente), algo que podría ser esperable si tenemos en cuenta que la saliva está compuesta por células sanguíneas, tema que comentaremos en el siguiente apartado.

Tras el desarrollo de modelos robustos que presentan errores bajos se han comenzado a explorar otros horizontes. Recientemente se ha puesto el foco más allá del material genómico autosómico desarrollándose modelos de predicción de la edad empleando marcadores presentes en el cromosoma Y. Vidaki et al. (286) desarrollaron un modelo de vectores de soporte compuesto por 19 posiciones CpG que presentaba un error de ± 8,46 años. Una de las notas más interesantes de este estudio es que no se observó una predicción peor en edades avanzadas en comparación con las observadas en individuos jóvenes, algo discordante con respecto a los modelos construidos con marcadores autosómicos. Recientemente Jiang et al. (287) construyeron un modelo empleando bosques aleatorios con 13 posiciones CpGs obteniendo un error de ± 5,73 años. Teniendo en cuenta la importancia del estudio del cromosoma Y en casuística forense, estos modelos presentan una aplicación muy interesante para el futuro del campo. Aunque su situación no sea comparable a la de los modelos con posiciones CpG autosómicas y se requieran aún más estudios, el desarrollo de este tipo de modelos podría posibilitar la aplicación de modelos de predicción en mezclas compuestas por dos individuos, siendo un contribuyente femenino y otro masculino.

Gracias a todos estos estudios, la predicción de la edad en sangre es una realidad aplicable, existiendo múltiples modelos que permiten obtener bajos errores de predicción que pueden reducir en gran medida el número de sospechosos en casos criminales. Pero la sangre no es el único tejido de interés forense, por tanto, se deben evaluar otros tejidos a fin de estar preparados ante la llegada de cualquier tipo de muestra.

c) Modelos de predicción de la edad en saliva y mucosa oral

Ante el éxito observado en los modelos de predicción de sangre, la atención comenzó a desviarse hacia otros tejidos de interés forense, entre los que se encuentran la saliva y la mucosa oral (hisopo bucal). La situación con estos tejidos no sería tan sencilla como en sangre, aflorando un problema asociado a la composición celular de los tejidos de interés. La heterogeneidad de la sangre demostró una consistencia en los patrones de metilación asociados a marcadores correlacionados con la edad que permitió su tratamiento como tejido homogéneo, pero al evaluar muestras de saliva e hisopo bucal la situación fue completamente distinta. Los tejidos que comparten la cavidad oral están compuestos principalmente por células sanguíneas (leucocitos) y células epiteliales (197), pero la complejidad emerge cuando se evalúa de forma independiente su presencia y proporción en los tejidos de interés. Al evaluar su composición, se observó que las muestras de saliva estaban predominantemente compuestas por células sanguíneas presentando una proporción reducida de células epiteliales (47,3%), mientras que la

población celular mayoritariamente representada en muestras de hisopo bucal eran células epiteliales (83,4%) (197). La convivencia de dos poblaciones celulares conlleva un problema asociado, la coexistencia de patrones de metilación diferentes en una misma muestra. Eipel et al. (195) evaluaron la presencia de leucocitos en hisopo observando un amplio rango de proporciones, entre el 12% y el 63%. Estos resultados mantenían la consistencia de otros estudios previos, centrados en el estudio de STRs, que habían observado proporciones de entre el 16% y 96% en saliva y entre 5% y 60% en hisopo bucal (288). Esta característica específica de estos tejidos ha condicionado en cierta medida, pero no impedido, la construcción de modelos de predicción de la edad. A continuación, en la Tabla 2, se presentan los modelos desarrollados con fines forenses más destacables asociados a estos tejidos.

Tabla 2. Resumen de los principales modelos de predicción de la edad en base a muestras de saliva e hisopo bucal. Se aporta la referencia de la publicación del modelo, año de publicación, genes en los que se encuentran los marcadores empleados, tejido analizado, metodología técnica utilizada y error obtenido. El número de genes no representa el número de marcadores empleado en la construcción del modelo.

Referencia	Año publicación	Genes empleados	Tejido	Metodología	Error (MAE)
Bocklandt et al. (289)	2011	*EDARADD* *NPTX2* *Tom1L1*	Saliva	Pirosecuenciación	± 5,2 años
Eipel et al. (195)	2016	*PDE4C* *ASPA* *ITGA2B*	Hisopo	Pirosecuenciación	± 4,3 años
Hong et al. (290)	2017	*SST* *CNGA3* *KLF14* *TSSK6* *TBR1* *SLC12A5*	Saliva	SNaPshot	± 3,13 años
Jung et al. (275)	2019	*ELOVL2* *FHL2* *KLF14* *C1orf132* *TRIM59*	Saliva Hisopo	SNaPshot	± 3,55 años ± 4,29 años
Schwender et al. (291)	2021	*PDE4C* *EDARADD* *KLF14*	Hisopo	Pirosecuenciación SNaPshot	± 5,11 años ± 5,16 años
Wozniack et al. (276)	2021	*PDE4C* *MIR29B2CHG* *ELOVL2* *KLF14* *EDARADD*	Hisopo	MiSeq®	± 3,7 años
Lee et al. (292)	2022	*SST* *KLF14* *TSSK6* *SLC12A5*	Saliva	PCR digital en gotas	± 3,3 años
Marcante et al. (293)	2024	*ELOVL2* *FHL2* *KLF14* *C1orf132* *TRIM59*	Hisopo	SNaPshot	± 3,5 años

Como se puede observar en la tabla 2, el primer modelo de predicción de edad forense se desarrolló en saliva en 2011 presentando un error de ± 5,2 años (289). Sería a partir de 2016 cuando la cavidad oral volvería a gozar de interés diferenciándose entre muestras de saliva e hisopo bucal. Los modelos de predicción de la edad desarrollados en base a estos tejidos

presentan una menor variedad, en comparación con los de sangre, en cuanto a tecnologías y metodologías estadísticas empleadas en su desarrollo. Esto no ha mermado sus resultados alcanzando con presteza el error de ± 3 años, presentando los modelos de saliva errores entre ± 3,13 y ± 5,2 años (275,289,290,292) y los de hisopo ± 3,5 y ± 5,16 años (195,275,276,291,293). Las tecnologías empleadas para el desarrollo de estos modelos han sido principalmente SNaPshot (275,290,291,293) y pirosecuenciación (195,289,291), comenzando recientemente su acercamiento a plataformas de MPS (276) y a nuevas tecnologías para el campo como la PCR digital en gotas (ddPCR) (292). En cuanto a los marcadores empleados se observa el uso de genes correlacionados con la edad en sangre como por ejemplo *ELOVL2*, *EDARADD*, *PDE4C* y *FHL2*. El uso de este tipo de posiciones se podría justificar por la presencia de leucocitos en la cavidad oral, formando parte de la composición celular tanto de muestras de saliva como de hisopo. Como se comentó anteriormente, la proporción de estas poblaciones es muy variable condicionando en cierta medida el desarrollo de estos modelos o los futuros planteamientos de la predicción de la edad en estos tejidos. Además, la realidad forense podría limitar la aplicabilidad de estos modelos a muestras desconocidas, ya que las proporciones celulares son muy variables entre individuos y no podemos estar seguros de qué tipo de tejidos estamos analizando o si el vestigio recuperado es una mezcla de ambos, como por ejemplo en una colilla. Este tipo de muestras, las colillas, han sido evaluadas mediante la predicción de la edad empleando un modelo construido en base a muestras de saliva, proporcionando un error de ± 7,65 años (294). Estos resultados refuerzan la necesidad de analizar los tejidos procedentes de la cavidad oral como un todo o de introducir factores asociados a la composición del tejido en los modelos de predicción de la edad. Con el objetivo de mejorar los modelos desarrollados se han planteado diferentes acercamientos que permitan reducir los errores de predicción obtenidos para estos tejidos. Eipel et al. (195), evaluó la introducción de marcadores específicos de población celular en los modelos de predicción de la edad, a fin de mejorar los resultados teniendo en cuenta el porcentaje de células epiteliales. En este caso los resultados observados con el modelo que combina marcadores de tipo celular y marcadores correlacionados con la edad no mejoran los errores obtenidos por estos últimos en solitario (± 4,66 años y ± 4,3 años respectivamente), siendo necesarios más estudios que permitan aclarar el potencial de este acercamiento en el estudio de muestras procedentes de la cavidad oral. Un acercamiento similar pero ligeramente diferente fue abordado por Hong et al. (290) en el desarrollo de su modelo de predicción para muestras de saliva. En este caso, en lugar de introducir un marcador específico de tipo celular se optó por uno de tejido, obteniéndose un menor error de predicción al introducirlo en el modelo de edad (± 3,2 años y ± 4,1 años con y sin marcador de tejido, respectivamente). La composición celular de las muestras de la cavidad oral es variable entre individuos, lo que puede ocasionar errores de predicción más elevados dependiendo de la muestra analizada. Aun así, la tendencia parece decantarse por un mayor porcentaje de células sanguíneas en saliva, algo que se ve reflejado al intentar aplicar modelos de predicción basados en muestras de sangre a muestras de saliva (± 3,3 años) e hisopo (± 14,6 años), presentando la saliva errores de predicción sustancialmente inferiores (180,195). Estas grandes diferencias no parecen estar asociadas a los marcadores seleccionados ya que, el reentrenamiento con muestras de hisopo bucal de modelos desarrollados con muestras de sangre han presentado errores de

entre ± 3-4 años (195,295), lo que parece indicar que el factor de mayor peso es la composición celular y los patrones de metilación derivados de ella. La presencia de leucocitos en muestras de saliva e hisopo abren la posibilidad a la construcción, con los mismos marcadores, de modelos combinados, algo que ya ha sido explorado obteniendo resultados muy prometedores al construir un modelo de predicción de la edad para muestras de sangre, saliva e hisopo bucal con un error de ± 3,84 años (275), pero que precisa de una mayor validación. Estos trabajos nos llevan a pensar que la construcción de un modelo combinado de aplicación forense es una realidad más cercana de lo que podíamos esperar y un acercamiento más correcto y aplicable a casuística forense.

Teniendo en cuenta lo comentado en este apartado, el futuro de los modelos de predicción de la cavidad oral podría estar asociado con la construcción de modelos multitejido. Por el momento, debería evaluarse la capacidad de los modelos existentes, o generarse nuevos modelos, para enfrentarse a muestra de saliva e hisopo, indistintamente, o a la mezcla de ambas, lo que podría representar una colilla en una investigación criminal. La aplicabilidad de estos modelos pasa por su robustez ante las diferentes composiciones celulares que pueden observarse tanto en muestras de saliva como de hisopo.

d) Modelos de predicción de la edad en semen

A parte de los tejidos previamente mencionados existe otro al que asiduamente los laboratorios forenses tienen que enfrentarse, las muestras de semen. El estudio de los patrones de metilación de ADN somático y germinal ha mostrado un panorama muy diferente entre estos dos materiales genéticos. Como se comentó anteriormente, el ADN somático presenta una hipometilación general e hipermetilación en posiciones específicas, pero estos patrones se ven invertidos en células germinales como los espermatozoides y células progenitoras (296,297). Esta tendencia discrepante ha imposibilitado una correcta evaluación de dichas muestras en relojes epigenéticos como el de Horvath (263), presentando muestras de semen evaluadas un error de ± 13,3 años (172). Esta situación se acrecienta sustancialmente cuando las muestras se analizan con modelos reducidos de uso forense, obteniéndose errores de ± 37,3 años con modelos de sangre compuestos por tres CpGs (298). Con estos resultados queda claro que los marcadores asociados con la edad en ADN somático no son aplicables a semen y *viceversa*, por tanto, se deben identificar marcadores específicos para este tejido. Teniendo en cuenta estas discrepancias tan marcadas es importante conocer la composición celular del tejido, ya que la contaminación con células somáticas podría ocasionar unos errores de predicción más elevados. La composición de muestras de semen ha sido evaluada identificándose que el 85% de células presentes son espermáticas. Si bien es cierto que el porcentaje restante del 15% puede parecer mucho, el 84% de este conjunto se ha asociado a células progenitoras (299). En la fracción restante se pueden encontrar células epiteliales y células sanguíneas, algo que debe tratarse con cuidado ya que estos porcentajes, al igual que en saliva e hisopo, podrían variar entre individuos. Esta composición y las diferencias tan abruptas entre tipos celulares podrían ser los motivos por los que los modelos de predicción desarrollados en base a muestras de semen presentan, generalmente, errores más elevados que los anteriores tejidos evaluados. A

continuación, en la Tabla 3, se recogen algunos de los modelos de predicción de la edad en base a muestras de semen más destacados.

Tabla 3. Resumen de los principales modelos de predicción de la edad en base a muestras de semen. Se aporta la referencia de la publicación del modelo, año de publicación, genes en los que se encuentran los marcadores empleados, metodología técnica utilizada y error obtenido. El número de genes no representa el número de marcadores empleado en la construcción del modelo.

Referencia	Año publicación	Genes empleados	Metodología	Error (MAE)
Lee et al. (172)	2015	*TTC7B* cg06979108 (sin gen) *NOX4*	SNaPshot	± 4,2 años
Li et al. (300)	2017	cg06979108 (sin gen)	Pirosecuenciación	± 4,05 años
Pisarek et al. (301)	2021	*SH2B2* *FOLH1B* *EXOC3* *IFITM2* *GALR2* *GALR2*	Pirosecuenciación	± 4,3 años
Heidegger et al. (264)	2022	*SYT7* *TUBB3* *SH2B2* *ARHGEF17* *EXOC3* *GALR2* *PPP2R2C* *TBX4* *PALM* *IFITM2* *NOX4* *TTC7B* *LOC401324*	MiSeq®	± 5,1 años

Los modelos construidos hasta el momento presentan unos errores comprendidos entre los ± 4-5 años (172,264,300–302), algo que solamente se ha reducido a ± 2,04 en un modelo desarrollado con 51 posiciones CpG (261), un número de marcadores mayor del habitual en forense. Por el momento queda esperar y ver cómo evoluciona el campo. Los conjuntos de entrenamiento y de validación empleados en los modelos actuales son relativamente pequeños ofreciendo la posibilidad de construir nuevos y mejores modelos. Sería conveniente evaluar la importancia de la composición celular y si la inclusión de marcadores específicos de tejido o de población celular podrían propiciar unas mejores predicciones. Por otro lado, debe plantearse la posibilidad de aplicar este tipo de modelos a muestras que presenten mezcla de dos individuos, siendo en muchos casos un contribuyente masculino y otro femenino, escenarios muy frecuentes en el ámbito forense. Ante estas situaciones Lee et al. (303) plantean una aproximación muy interesante, el uso de marcadores correlacionados con la edad específicos de cromosoma Y. Si bien es cierto que por el momento los errores obtenidos son más elevados de lo habitual, entre ± 5-7 años para las distintas aproximaciones estadísticas evaluadas, las posibilidades de estos modelos son muy esperanzadoras consiguiendo franquear la imposible interpretación de mezclas de metilación de ADN actual, escenario en el que es habitual encontrar muestras de semen en rutina forense.

e) Modelos de predicción de la edad en restos cadavéricos

Cuando pensamos en genética forense existen unos últimos tejidos por comentar que nos vienen con facilidad a la mente, los restos cadavéricos. Estos tejidos se van a tratar como un conjunto tanto por el limitado número de análisis existente como por su utilidad como posibles vestigios biológicos para la predicción de la edad. Entre estos tejidos, que presentan un mayor desafío que los anteriores por las condiciones en las que se encuentran, podemos encontrar huesos, dientes, pelos, uñas o músculo. En los últimos años se han identificado marcadores epigenéticos correlacionados con la edad en muestras cadavéricas y se han comenzado a desarrollar los primeros modelos de predicción, presentándose un resumen de los más destacados en la Tabla 4.

Tabla 4. Resumen de los principales modelos de predicción de la edad en base a muestras de restos cadavéricos. Se aporta la referencia de la publicación del modelo, año de publicación, genes en los que se encuentran los marcadores empleados, metodología técnica utilizada, tejido y error obtenido. El número de genes no representa el número de marcadores empleado en la construcción del modelo.

Referencia	Año publicación	Genes empleados	Metodología	Tejido	Error (MAE)
Bekaert et al. (282)	2015	*ASPA* *PDE4C* *ELOVL* *EDARADD*	Pirosecuenciación	Diente	± 4,86 años
Marquez-Ruiz et al. (304)	2020	*ELOVL2* *PDE4C*	Pirosecuenciación	Diente	± 5,08 años
Wozniak et al. (276)	2021	*ELOVL2* *KLF14* *PDE4C* *ASPA*	MiSeq®	Hueso	± 3,4 años
Hao et al. (305)	2021	*LAG3* *SCGN* *ELOVL2* *KLF14* *C1orf132* *SLC12A5* *GRIA2* *PDE4C*	SNaPshot	Pelo	± 3,68 años
Zapico et al. (306)	2021	*ELOVL2* *NPTX2* *KLF14* *SCGN* *FHL2*	Pirosecuenciación	Diente	± 1,5 años
Fokias et al. (307)	2023	*ASPA* *EDARADD* *PDE4C* *ELOVL2*	Pirosecuenciación	Uñas	± 4,82 años

El desarrollo de estos modelos puede ser de gran importancia en diversos escenarios forenses en los que las predicciones de edad en antropología presentan unos errores más elevados, como ya hemos visto en apartados anteriores. En un primer momento, la correlación con edad en huesos se evaluó teniendo en cuenta marcadores de sangre. Si bien es cierto que los tejidos presentaban diferencias en sus patrones de metilación, algunos de los marcadores que han presentado una mayor correlación con la edad en sangre la presentan también en huesos (308,309). Estas similitudes quedan patentes al obtener Lee et al. (309) errores de ± 6,4 años al evaluar muestras de huesos empleando el modelo de sangre y piel desarrollado por Horvath

(310). Actualmente el mejor modelo de predicción desarrollado para muestras de hueso es el construido por Wozniak et al. (276) empleando seis posiciones CpG. Dicho modelo presenta un error de ± 3,4 años y fue evaluado con muestras de sangre, cartílago y musculo, presentando unos errores de ± 4,9, ± 25,8 y ± 13,7 años respectivamente. Como se puede observar, los resultados obtenidos para muestras de sangre con un modelo de huesos destacan las similitudes de los patrones de metilación de ambos tejidos. La construcción de modelos multitejido que combinen hueso y sangre ha sido abordada, incluyendo en la ecuación otro vestigio biológico de interés forense, las piezas dentales. Dias et al. (311) construyeron un modelo combinando estos tejidos y obtuvieron un error de predicción de ± 6,06 años. La aplicabilidad de este modelo debe ser evaluada detenidamente teniendo en cuenta las condiciones extremas en las que, en algunas ocasiones, se encuentran los huesos analizados en genética forense. La sensibilidad y resistencia de los patrones de metilación ante la degradación son factores fundamentales que deben ser analizados cuidadosamente a fin de introducir estos modelos en casuística forense.

El estudio de los dientes para la predicción de la edad también partió de su comparación con patrones de metilación en sangre. Bekaert et al. (282) evaluó un conjunto de muestras de diente en su modelo desarrollado en base a muestras de sangre, obteniendo unos errores de predicción de ± 4,86 años. Otros estudios centrarían sus esfuerzos en la evaluación de las diferentes capas del diente (dentina, cemento y pulpa) analizando la cantidad de ADN obtenido a partir de cada una de ellas y desarrollando modelos específicos para cada uno de esos estratos (312). Al evaluar los patrones de metilación de dentina, cemento y pulpa no se observaron diferencias en media y varianza entre ellos, pero sí que se obtuvieron diferencias en los errores obtenidos en los modelos desarrollados (± 7,0, ± 2,45 y ± 2,25 años, respectivamente). Estos resultados deben ser evaluados cuidadosamente y no puede hacerse una comparación directa asociada a la capa analizada. El alto error obtenido para dentina, en comparación con cemento y pulpa, puede estar asociado al uso de un número reducido de marcadores, cinco frente a 13 CpGs. De todas formas, sería interesante realizar más estudios desde esta perspectiva ya que parte de las discrepancias obtenidas en los errores de predicción podrían ser explicadas por la diferente composición celular de las capas del diente. Mientras que la dentina presenta un alto porcentaje de minerales, el cemento y la pulpa tienen una composición más variada con células específicas del diente (odontoblastos, cementocitos y cementoblastos) y riego sanguíneo procedente de capilares y venas (313), lo que podría explicar la similitud de los valores de metilación entre dientes y sangre (312). Por último, otro acercamiento interesante fue el estudio de la combinación, para la predicción de la edad en muestras de diente, de dos procesos asociados al envejecimiento, la metilación del ADN y el acortamiento telomérico (304). La construcción de un modelo basado exclusivamente en metilación de ADN y otro en base al acortamiento de los telómeros demostró que la metilación del ADN permitía obtener una mejor predicción de la edad, proporcionando un menor error de predicción (± 5,08 y ± 6,89 años, respectivamente). Al evaluar estas variables de forma combinada, el error obtenido por el modelo presenta una mejoría ínfima, ± 5,04 años, en comparación con el modelo basado en metilación del ADN. Por el momento, los modelos desarrollados con piezas dentarias están construidos con números de muestras limitados y, aunque se hayan obtenido errores de

predicción bajos, han sido escasamente validados por lo que su aplicación forense está lejos de ser una realidad.

Los huesos y los dientes han sido los restos cadavéricos más estudiados, pero en los últimos años los horizontes se han ampliado al estudio del cabello, uñas y músculo. Se han identificado patrones de metilación correlacionados con la edad en cabello (314), pero hay ciertas condiciones que deberían evaluarse. Cabe destacar que el ciclo de crecimiento del cabello no está sincronizado, siendo posible analizar pelos en distintas etapas del ciclo lo que podría provocar, algo que se ha observado en cabras, diferencias en los patrones de metilación observados (315). Esto podría extenderse a las diferencias observadas en el pelo del cuerpo con respecto al del cuero cabelludo o si el encanecimiento del cabello y, por tanto, la pérdida de actividad de los melanocitos puede derivar en patrones de metilación diferenciados. Estas incógnitas que aún deben ser resueltas no han frenado la creación de un modelo de predicción cuyo error fue ± 3,68 años (305). El modelo fue evaluado en muestras de pelo de otras partes del cuerpo diferentes de la cabeza y en pelo de animal, un contaminante que, aunque fácilmente detectable mediante un estudio morfológico con microscopio, podría llevar a error de no comprobarse. Teniendo en cuenta los resultados obtenidos parece que estos factores no afectan al modelo, aunque debe destacarse la detección de picos inespecíficos en pelos de conejo y gato. Los resultados parecen prometedores pero una validación con un mayor número de muestras sería necesaria para discernir con certeza estas afirmaciones. Por otro lado, las uñas han sido también evaluadas como un posible vestigio a raíz del cual se podría obtener una predicción de la edad, pero se detectó un problema que debe ser abordado con cuidado. Para la construcción del modelo se tomaron muestras de uñas de las cuatro extremidades y, para alguna de las posiciones CpG analizadas, se observó una gran variabilidad en los patrones de metilación obtenidos como consecuencia de la localización de la toma de muestra (316). Esto llevó a los autores a la construcción un modelo conjunto con todas las muestras recogidas independientemente de su localización y modelos de predicción para cada una de las extremidades, obteniéndose errores de predicción entre ± 6,09 y ± 7,91 años. La posterior identificación de nuevos marcadores y la reconstrucción del modelo permitió reducir el error obtenido hasta los ± 4,82 años (307). Por último, el tejido muscular ha comenzado a llamar la atención del campo como un posible candidato para engrosar la lista de tejidos para el estudio de restos cadavéricos. A pesar de que existen relojes epigenéticos basados en este tejido que presentan errores cercanos a los ± 4 años (317), su incursión en el ámbito forense y la creación, por tanto, de relojes de edad cronológica aún es una tarea pendiente. Por el momento solo se ha llevado a cabo un tímido acercamiento a este tejido, comprobando la robustez de éste y otros modelos con muestras *post mortem* (318).

Como se puede observar, los modelos basados en muestras de restos cadavéricos se encuentran en un estado relativamente inmaduro, pero crecen con gran rapidez gracias a los conocimientos aportados al campo durante el estudio de otros tejidos. La realidad es que la obtención de algunas de estas muestras es una ardua tarea que en muchas ocasiones se escapa a nuestro control, pero las ganas de generar nuevo conocimiento por parte de la comunidad forense no dejan ir ninguna oportunidad interesante que pasa frente a ella. Durante un tiempo,

la sangre fue el tejido más estudiado, como si ningún otro existiese, su disponibilidad y los resultados observados opacaron temporalmente el estudio de otros vestigios, pero con el tiempo nos permitimos dejarlos brillar con luz propia, estando actualmente la atención más repartida. Como dijo una vez el gran escrito de fantasía John Ronald Reuel Tolkien, *"La luz de la luna ahoga a todas las estrellas salvo a las más brillantes"*, veremos pues qué estrellas seguimos en el futuro y qué descubrimos sobre ellas.

1.4.3.6 Inferencia de estilos de vida

En el ámbito forense se busca activamente formas de proporcionar mayor cantidad de información, con el objetivo de reducir el número de sospechosos y facilitar la identificación del donante de una muestra biológica recogida en la escena de un crimen. La metilación del ADN ya ha mostrado su capacidad para proporcionar información relevante en relación a un vestigio desconocido, pero aún quedan factores por evaluar en relación con las variables que tienen efecto sobre esta marca epigenética. Estas características son clave en el entendimiento de este biomarcador y en la identificación de nuevas aplicaciones tanto clínicas como forenses, siendo producida la mayor contribución a la variación de la metilación del ADN por influencias del ambiente (319). El ambiente no hace referencia sólo a lo que nos rodea, en el contexto que nos ocupa no se define como vivir en una zona costera o en una zona de montaña, hace referencia a como interaccionamos o como interacciona con nosotros el ambiente. Es cierto que parece un contexto amplio, pero debemos centrar nuestro objetivo en las rutinas cotidianas que tienen que ver con hábitos de vida. Por tanto, la metilación del ADN y los estilos de vida en el ámbito forense se centrarán, por ahora, en intentar definir si el donante de una muestra consume alguna droga o no, si hace ejercicio físico, si lleva una dieta saludable o si padece alguna enfermedad. Características que llegado cierto momento podrían ser de utilidad a los investigadores de un crimen. Cabe destacar que estos hábitos presentan complejas interacciones a nivel biológico y su predicción, en algunos casos, es desafiante. Los comienzos no son fáciles, pero es emocionante ver los primeros pasos y el establecimiento de nuevas aplicaciones.

a) Tabaco

El consumo de tabaco es un hábito que comenzó a globalizarse en 1492, no por su invención en aquella época, sino porque se observó, por primera vez, como los indígenas cubanos fumaban unas hojas secas que poco después serían nombradas comúnmente como tabaco. A lo largo de los años, la percepción social sobre el consumo de esta sustancia variaría pasando por etapas muy contrarias. Al principio, durante mucho tiempo, se vería como una práctica moderna en la época, elegante y asociada a cierto estatus y distinción social. Actualmente la historia es muy diferente y recorremos un camino, lentamente, de estigmatización y rechazo. Esta nueva tendencia social hacia el tabaco no parece repercutir en su consumo a nivel poblacional. Según los datos publicados en 2023 por el Instituto Nacional de Estadística (320) (INE) y el Observatorio Español de las Drogas y las Adicciones (321) (OEDE) se ha observado un incremento de un 13% en la percepción de riesgo desde 1997, derivado de las políticas de concienciación. Aún con todo, la prevalencia de consumo de tabaco se ha mantenido más o menos constante a lo largo de los años siendo de un 34,9% en 1997 y de un 33,1% en 2020.

También es importante destacar que el consumo medio de cigarrillos al día se sitúa en 11,9, una cifra que, a ojos de un no fumador, parece ciertamente elevada. Con estos datos podemos extrapolar que es difícil que se produzca un cambio abrupto en la prevalencia de consumo y que el elevado número de cigarrillos consumidos puede derivar en diferencias más marcadas en los patrones de metilación entre no fumadores y fumadores. Pero ¿cómo se distribuye este consumo en la población? La respuesta a esta pregunta proporcionará la informatividad de la aplicación, identificando al grupo minoritario que potencialmente podría reducir el número de sospechosos en un acto criminal. Según los datos publicados por el INE en el día mundial sin tabaco de 2023 (320), el porcentaje de no fumadores en España es del 77,9%, siendo el 55,9% no fumadores y el 22% exfumadores. Por otro lado, el porcentaje de fumadores es de 22,1% (19,4% diario y 2,3% ocasional). Si extendemos esta búsqueda a Europa, recurriendo a los datos de Eurostat de 2019 (322), el porcentaje de fumadores diarios es de 18,4%, un valor muy similar al de España. Gracias a estos estudios podemos entender el ambiente y definir el interés a nivel clínico y forense del estudio de este estilo de vida. En el caso que nos compete, el campo forense, la correcta predicción del grupo minoritario (fumadores) podría proporcionar información relevante en ciertas investigaciones.

El consumo y efecto de esta sustancia en la salud de sus consumidores ha sido ampliamente estudiada, generándose, gracias a la metilación del ADN, una nueva capa de información (323). En los últimos años se han publicado diversos estudios que evalúan las diferencias de metilación entre no fumadores y fumadores, identificado una tendencia a la hipometilación en consumidores en comparación con no consumidores (324,325). Estas observaciones ofrecen una perspectiva de seguimiento muy interesante para el ámbito clínico. La asociación de la metilación del ADN con la salud provoca el surgimiento de múltiples estudios que evalúan los efectos del consumo de tabaco en diversos desordenes. El cáncer fue uno de los primeros objetivos de estudio (326), un movimiento lógico al ser el cáncer de pulmón una de las diez principales causas de muerte en el mundo según la Organización Mundial de la Salud (OMS) (327). Este estudio fue el primero de muchos que siguieron su estela, correlacionando el efecto del consumo de tabaco sobre la metilación y diversas enfermedades, dolencias o condiciones de la salud: efecto en el sistema neuroendocrino (328), desarrollo de fragilidad en edades avanzadas (329), consumo materno prenatal y saludo infantil (330), hipertensión (331) y enfermedades cardiovasculares (332), entre otras. Todos estos estudios que describen asociaciones llevan al ámbito clínico por el mismo camino que a continuación se describirá en forense, siendo la siguiente parada el desarrollo de modelos de predicción. En el ámbito clínico dichos modelos tienen como objetivo identificar el riesgo de padecer una enfermedad o identificar una enfermedad en etapas tempranas del desarrollo. Actualmente hay publicados estudios que permiten clasificar entre tejido normal y canceroso a partir de células bucales (333), predecir el riesgo cardiovascular de fumadores (334) y el daño pulmonar en fumadores y exfumadores (335). Si bien es cierto que aún son necesarios un mayor número de estudios, estos trabajos nos muestran el gran potencial que esconde el estudio de la metilación del ADN.

Ahora bien, debemos desviar nuestra atención al campo que nos compete. Para comprender el potencial del estudio de la relación del consumo de tabaco y la metilación del ADN en el

ámbito forense, debemos tener en cuenta la visión general de la distribución en la población comentada anteriormente. En forense, la información que puede generarse está relacionada con la clasificación de una muestra en una de las categorías de la población, definidas en este contexto como no fumadores, exfumadores o fumadores. Lo primero que debemos plantearnos es cómo se comporta la metilación del ADN en base a estas categorías. Como se comentó en el párrafo anterior, se identificaron grandes diferencias entre los grupos extremos, no fumadores y fumadores, lo que llevó a la generación de modelos de clasificación que solo tenían en cuenta estas categorías. Dichos modelos, basados en regresión logística binomial, presentaron áreas bajo la curva (AUCs) superiores a 0,90 (336–340), un nivel de clasificación muy elevado, pero ¿a qué coste? Construir un modelo evaluando solo dos categorías genera un modelo sesgado, un modelo que no representa a la población general y que, por consiguiente, no es aplicable. Por tanto, ¿cómo se comportan los patrones de metilación del grupo intermedio? Uno de los descubrimiento más interesantes fue la reversibilidad o irreversibilidad de ciertas posiciones ante el cese de consumo (341–344). Gracias a estos estudios se identificaron posiciones que en exfumadores recuperaban los valores de metilación a niveles de no fumadores entre 0 y 35 años después del cese, y posiciones que mantenían patrones de metilación de fumador 35 años después. Estos descubrimientos han sido de gran interés a nivel clínico porque, como destaca Guida et al. (342), las posiciones reversibles imitan los perfiles de riesgo observados tras el cese de consumo, relacionando la reversibilidad de las posiciones y la reducción del riesgo de, por ejemplo, cáncer de pulmón. A nivel forense, estos patrones nos aportan una información clave a la hora de evaluar los modelos de clasificación generados. Por tanto, los exfumadores conforman un grupo más heterogéneo y difícil de predecir, condicionado por la reversibilidad, el tiempo desde el cese del consumo y la intensidad de consumo (345). Ante esta situación se plantean dos posibles soluciones que han sido evaluadas en los diversos modelos generados a lo largo de los años; agrupar dos de las tres categorías, generando un modelo de clasificación binomial, o construir un modelo multinomial y clasificar las tres categorías por separado. Puede que la segunda opción sea la más adecuada, pero su ejecución y aplicación son más difíciles de lo que parece. La complejidad de los patrones de metilación observados en exfumadores provoca que la separación completa de las categorías analizadas sea un reto aún pendiente de solventar. Esto, en los modelos multinomiales, se traduce en una correcta predicción de los grupos extremos y en una predicción errónea del grupo intermedio, los ejemplos más claros se presentan en los trabajos de Alghanim et al. (346) y Maas et al. (347). Ambos trabajos, empleando metodologías (pirosecuenciación y MPS, respectivamente) y marcadores (4 CpGs y 13 CpGs, respectivamente) diferentes presentan tendencias de clasificación similares. En ambos trabajos se observa como la capacidad de clasificación de los exfumadores se ve mermada con respecto a los grupos extremos. Para entender estos resultados debemos preguntarnos en qué grupos se clasifican estos exfumadores, si siguen una tendencia o si están repartidos indistintamente entre no fumadores y fumadores. Alghanim et al. construyen modelos binomiales, evaluando las cuatro CpGs seleccionadas individualmente, enfrentando las tres categorías entre sí; no fumadores *vs* fumadores (AUCs: 0,95-0,98), no fumadores *vs* exfumadores (AUCs: 0,72-0,75) y exfumadores *vs* fumadores (AUCs: 0,86-0,94). A la vista de sus resultados se observó una mayor dificultad a la hora de distinguir no fumadores y

exfumadores llevándonos a pensar que la reversibilidad de los patrones de metilación tiene un gran peso en este grupo intermedio. La idea de agrupar estas dos categorías de no consumidores ha sido evaluada (340,345,347,348) presentando todos ellos un alto grado de clasificación (AUCs: 0,90-0,99).

La predicción de estado de fumador ha avanzado en los últimos años a pasos agigantados con tres pilares fundamentales sobre los que se sustenta. Las diferencias observadas entre los patrones de metilación de no fumadores y fumadores han permitido, como se ha destacado anteriormente, diferenciar correctamente entre dichas categorías. Por otro lado, la categoría intermedia, exfumadores, ha demostrado una tendencia de reversibilidad y agrupación con los no fumadores que permite generar modelos de clasificación representativos de la población, haciendo que el grupo no sea inclasificable con los marcadores y medios actuales. El último pilar, no comentado hasta el momento, es la presencia de marcadores que presenten un alto poder de discriminación. En el caso de consumo de tabaco sería justo destacar una posición CpG en concreto, cg05575921, presente en el gen *AHRR*. Esta posición no solo está presente en la mayor parte de los modelos publicados (337–340,345,347–349), sino que también ha demostrado su potencial de forma individual alcanzando valores de AUCs entre 0,96-0,99 al clasificar no fumadores frente a fumadores (337,338,348). Aunque los modelos de clasificación de estado de fumador presentados hasta la fecha presenten un alto poder de clasificación, existen ciertos inconvenientes y desventajas que limitan en gran medida la aplicabilidad de dichos modelos. Hoy en día, los procesos de validación forense aplicados a los modelos desarrollados son escasos, un paso de gran importancia en favor de una futura implementación. Esta validación es crucial en modelos que sustenten la mayor parte de su poder de discriminación en un solo marcador, como podría ser el caso de *AHRR*, cuya pérdida al analizar una muestra pusiese en grave peligro los resultados obtenidos. En relación con su aplicabilidad se presenta un inconveniente recurrente al emplear metilación del ADN, el tejido de estudio. El escenario que se nos presenta en esta parte está monopolizado por los modelos basados en muestras de sangre, el único acercamiento a otros tejidos fue la aplicación de dos modelos desarrollados en sangre a muestras de saliva (346,348). En este aspecto es necesario un mayor número de estudios con el objetivo de identificar marcadores de estado de fumador que permitan desarrollar modelos específicos de tejidos de interés forense. Estas trabas no son lapidarias y el estudio de la relación de la metilación del ADN con el consumo de tabaco sigue avanzando de forma muy prometedora. A parte de los modelos desarrollados, se llevan a cabo estudios que pretenden predecir características más concretas asociadas a exfumadores y fumadores. Hemos visto que los exfumadores son un grupo difícil de predecir por separado, por ello, en paralelo con los modelos, se llevan a cabo estudios de características asociadas a este grupo que podrían aportar un mayor grado de información, como por ejemplo duración y tiempo desde cese de consumo (347,350). Por el momento la introducción de estas covariables en los modelos desarrollados no ha sido aplicada con éxito, obteniéndose correlaciones no lo suficientemente elevadas como para ser informativas (R^2: 0,47-0,51 y AUCs: 0,78-0,79). La búsqueda de marcadores relacionados con estas características podría ser clave en su futura implementación. A su vez, de forma adicional al modelo de clasificación se comenzó a evaluar

la posibilidad de aportar más información sobre aquellos individuos clasificados como fumadores, por ejemplo, el número de paquetes de cigarrillos consumidos al año (≥ 10 vs < 10 o ≥ 15 vs <15 paquetes al años) (347), los años de consumo (350) o años desde cese de consumo (351). Si bien es cierto que las correlaciones observadas son ligeramente más elevadas que en el caso anterior (AUC: 0,81-0,85 y R^2: 0,64, respectivamente) aún no han sido aplicadas de forma efectiva en ningún modelo de predicción.

El consumo de tabaco es, sin lugar a duda, el estilo de vida más estudiado a nivel forense y, por el momento, que mejores resultados ha presentado. Aún queda mucho que mejorar, pero los cimientos generados son lo suficientemente sólidos como para considerar su aplicabilidad bastante más próxima y realista que la de los estilos de vida que comentaremos a continuación.

b) Alcohol

Las bebidas alcohólicas llevan acompañado al ser humanos desde los albores del antiguo Egipto, donde las bebidas fermentadas gozaban de una alta consideración. En aquellas épocas, donde el agua de ríos y pozos podía contener enfermedades mortales, la cerveza, gracias a su contenido en alcohol, era considerada la bebida más sana, siendo disfrutada, como el *Grand Prix*, por niños y mayores por igual. Su popularización surgió en diversas civilizaciones y los griegos en su literatura comenzaron a advertir sobre su exceso de consumo. En el siglo XVIII, en Reino Unido, se aprobó una ley que fomentaba la creación de bebidas espirituosas, provocando una generalización del alcoholismo. A partir del siglo XIX la percepción comenzó a cambiar, las campañas antialcohólicas empezaron a surgir, llegando a instaurarse leyes de prohibición total como la famosa "ley seca", aplicada en Estados Unidos. Este tipo de normativas no sobrevivieron mucho y hoy en día convivimos con dichas bebidas en una aceptación social medida por el control de consumo de cada individuo. Lamentablemente, el autocontrol no es suficiente y al año fallecen millones de personas en el mundo por culpa del consumo de bebidas alcohólicas. Esto ha hecho que se convierta en un tema de interés para la investigación biomédica, iniciándose una búsqueda de marcadores genéticos asociados con el abuso del alcohol (352–354) o la evaluación de su heredabilidad (355) con el objetivo de convertirse en la base farmacológica de su tratamiento (356). Gracias este tipo de estudios, el alcohol empezó a definirse como factor de riesgo para múltiples enfermedades: hipertensión, cirrosis de hígado, pancreatitis crónica, enfermedades coronarias de corazón y cáncer bucal, de esófago y de laringe (357). Este efecto tan generalizado en nuestro organismo dificultó el estudio de los mecanismos subyacentes y las respuestas empezaron a buscarse en otro lugar, el epigenoma. Los estudios epigenéticos llevados a cabo fueron diversos y variados, conformando una limitación que dificultaría el diseño de modelos predictivos o índices de riesgo. Los diseños de los estudios eran muy diferentes y los descubrimientos no eran replicables (358). ¿En qué se tradujo esto? Si bien es cierto que se identificaron diferencias entre no bebedores y bebedores, las tendencias identificadas en los patrones de metilación variaban, identificándose patrones generales de hipermetilación (359–361), patrones generales de hipometilación (362–364) o ambos simultáneamente (365). Pero esto solamente ralentizaría un poco el avance. Si bien es cierto que no se identificaron marcadores hegemónicos como el gen *AHRR* en el consumo de

tabaco, existían diferencias que daban pie a la diferenciación entre categorías. A su vez algunos de estos estudios permitieron identificar características importantes de la metilación del ADN asociadas al consumo de alcohol, como por ejemplo su gran impacto en el epigenoma en general (364), la reversibilidad de los patrones de metilación tras el cese de consumo (364,366) o la observación de cambios en el epigenoma tras los seis primeros meses de consumo (360). Aún con los inconvenientes encontrados por el camino, los modelos de clasificación empezarían a surgir poco a poco siendo de gran interés identificar a los bebedores altos tanto para el ámbito clínico como forense.

A nivel sanitario, la identificación de personas con elevado consumo de alcohol ha generado interés con el objetivo de mejora de diagnósticos. Algo lógico teniendo en cuenta los graves efectos de salud que conlleva un consumo irresponsable de dicha sustancia. Los primeros modelos generados con este fin estaban basados en marcadores convencionales, las proteínas, aunque presentaban, en sus inicios, valores de AUC comprendidos entre 0,21 y 0,67. Estas predicciones fueron mejorando con el tiempo al identificarse nuevas proteínas, alcanzando valores de entre 0,73 y 0,86 (367), pero los resultados dejaban margen de mejora, no solo por sus valores de AUC, si no también, porque sólo permitían clasificar entre no bebedores y bebedores altos. La epigenética coge el testigo y el campo forense, viendo las diferencias observadas en el epigenoma, comenzó a interesarse por la informatividad de la predicción de este estilo de vida. Identificar a un individuo como bebedor alto podría reducir en gran medida el número de sospechosos en una investigación criminal. Teniendo en cuenta las estadísticas nacionales, este grupo de bebedores comprende un pequeño porcentaje de la población, con un prevalencia de consumo diario del 9% (321) y con un 19,7% de los hombres y un 5.9% de las mujeres consumiendo diariamente bebidas alcohólicas (368). Pero en este caso las cosas no fueron tan rodadas como con tabaco a la hora de construir los modelos de clasificación, observándose valores de AUC que variaban entre 0,55 y 0,83 según los grupos de clasificación analizados y el número de marcadores empleados (340,345,369). ¿Cuál o cuáles podrían ser las causas de estos resultados tan dispares? Uno de los primeros problemas lo encontramos en la capa más superficial, la información de consumo de los participantes. Las personas suelen infraestimar el número de bebidas alcohólicas que consume e incluso, en ocasiones, dudan si considerar consumo de alcohol el tomarse un par de cervezas a la semana. Este problema se ha comentado en la bibliografía (366,369), condicionando en gran medida los datos recogidos y por ende, la correcta clasificación de los mismos. Otro problema relacionado con esta información es la agrupación de los participantes en los distintos grupos de clasificación. En este aspecto se presenta una mayor diversidad de grados de consumo en comparación con tabaco. Definir qué se considera bebedor alto o no bebedor puede ser sencillo, pero el rango intermedio es más disperso clasificándose en algunas ocasiones como un grupo conjunto o fragmentándolo en grupos más pequeños en otras. A su vez, los ex bebedores, aún con la reversibilidad observada en los marcadores epigenéticos, muestran patrones de metilación a medio camino entre bebedores y no bebedores (364), pudiendo solaparse con bebedores de bajo consumo. Si esto ya empieza a sonar complicado existe un aspecto más a tener en cuenta, el sexo del individuo en este estilo de vida cobra gran importancia. Se ha demostrado una diferente

sensibilidad y susceptibilidad asociada al metabolismo con respecto al consumo de alcohol en hombres y mujeres (370,371), lo que repercute en una diferencia en la cantidad de consumo sesgada por el sexo. Esto provoca que la clasificación de los individuos en las diferentes categorías definidas está condicionada por la cantidad de consumo y el sexo del individuo. Dejando atrás los problemas de clasificación, los propios patrones de metilación han supuesto un reto, como se comentaba anteriormente en este mismo apartado. Se ha observado que el consumo de alcohol tiene un gran impacto en el epigenoma general, lo que podría llevar a seleccionar marcadores que estén influenciados por otras variables. Si bien es cierto que esto nos lleva a pensar que hay mucha variación, la realidad es que las diferencias observadas entre no bebedores y bebedores altos son menores de un 10% (366). A su vez, dentro de estas diferencias, las más pronunciadas se identificaron en mujeres, no fumadores y en individuos con baja predisposición genética al abuso de alcohol (364). Con este escenario ante nosotros podemos comprender lo desafiante que puede llegar a ser la clasificación del estado de bebedor. Los nuevos modelos diseñados deben ser cuidadosos con la agrupación de los individuos, la evaluación de posibles covariables en los marcadores seleccionados y una construcción adecuada de los modelos de clasificación en base a los grupos definidos. A continuación, comentaremos estilos de vida que por el momento han tenido poco impacto en el ámbito forense, siendo su crecimiento derivado del interés clínico que suscitan. De todas formas, merecen un hueco en este apartado ya que se plantean como el futuro de esta aplicación.

c) Drogas duras

La historia del consumo de drogas comienza antes del 3000 a. C. cuando se empleaban como analgésicos, en algunas partes del mundo, el cáñamo, las hojas de coca o el peyote. Estas sustancias se consumían de forma localizada y su uso se fue extendiendo con los intercambios culturales y poblacionales que ha vivido nuestra civilización a lo largo de los siglos. Los efectos perjudiciales de estas sustancias han sido un punto de interés a nivel mundial en el ámbito sanitario, y la investigación sobre sus efectos y adicción se ha desarrollado a lo largo de los años. Gracias a estos estudios se ha confirmado que la adicción a las drogas se produce debido a la plasticidad neuronal (372), la capacidad del sistema nervioso de modificar su estado en función de las condiciones del medio. El consumo de drogas suele comenzar como algo esporádico, recreativo, siendo su paso a abuso crónico y adictivo producido por adaptaciones neuroplásticas en los circuitos cerebrales asociados al sistema de procesamientos de recompensas, creando de una relación de consumo-recompensa que se activa mediante centros de dopamina (373). El estudio de la plasticidad neuronal ha revelado que el epigenoma es un agente de mediación de gran importancia en el tipo de neuroplasticidad, la neuroplasticidad persistente, que se observa en la adicción a las drogas (374,375). Ante los investigadores surge una nueva pieza del rompecabezas, un nuevo factor de gran importancia en la interacción del organismo con el ambiente, propiciando el estudio del biomarcador que podría aportar una explicación al proceso biológico de la adicción a las drogas.

El campo clínico, una vez más, ha sido el ámbito que ha encabezado el estudio de la relación de la metilación del ADN y el consumo de drogas duras. Con el paso de los años y el

avance de las tecnologías comienzan a realizarse estudios de epigenoma completo, lo que permitirá avanzar de forma sustancial tanto en el ámbito clínico como forense. De las diversas drogas existentes, una de las más estudiadas es el cannabis, cuyo consumo se ha generalizado con el paso de los años convirtiéndose en una preocupación a nivel mundial. Aun cuando el consumo de esta sustancia se ha asociado a procesos neurodegenerativos o enfermedades mentales (376–379), su percepción de riesgo en la población es menor que la de tabaco. En el caso de España, según las estadísticas recogidas por el OEDA (321), la percepción de riesgo para el consumo diario de tabaco es del 92,1%, mientras que, en el caso del cannabis, para el consumo semanal es del 83,8%, descendiendo al 63,7% cuando es mensual. Estas cifras se han visto reflejadas en la prevalencia de consumo recurrente siendo un 8,6% en 2022, casi el doble que en 1997. Estos porcentajes hacen que, para el ámbito forense, la construcción de modelos de clasificación entre consumidores y no consumidores puedan conferir la posibilidad de aportar información altamente relevante sobre el donante de una muestra. Teniendo en cuenta los tejidos de interés forense, por el momento los estudios publicados se centran principalmente en sangre (380–384), observándose diferencias de metilación en zonas específicas, como por ejemplo el promotor del gen *CNR1* que codifica un receptor de cannabinoides, o generales analizando el epigenoma en estudios de asociación de genoma completo. Aunque la sangre es el tejido más evaluado, también se ha estudiado el efecto del consumo de cannabis en semen (385). El estudio de este tejido ha mostrado no solo diferencias en los patrones de metilación entre consumidores y no consumidores, sino que también, una menor concentración de esperma en consumidores. Estos trabajos no solo han aportado posiciones de interés, sino que también han identificado características de gran importancia que deben tenerse en cuenta a la hora de desarrollar modelos de clasificación. El consumo simultáneo de tabaco y cannabis es algo común, pero ¿cómo se ve reflejado el efecto de estas variables en el epigenoma? Esta pregunta tiene mucho peso a la hora de generar modelos de clasificación ya que el co-consumo de estas sustancias puede derivar en errores de asignación. Se ha demostrado que el efecto del tabaco sobre el epigenoma es mayor que el del cannabis, pudiendo enmascarar los patrones de metilación de este último (381). A su vez, la posición CpG de referencia para la clasificación de consumo de tabaco (cg05575921 en *AHRR*) se ve afectada por el consumo de cannabis, siendo el nivel de metilación menor para aquellos individuos que consumen tanto tabaco como cannabis (386), efecto que podría llevar a la clasificación errónea de un fumador. Por otro lado, se han identificado posiciones asociadas al consumo reciente o al consumo acumulativo de cannabis (383), siendo de gran interés para la construcción de modelos más robustos y precisos. Esto ha propiciado, aunque es preciso realizar más estudios, la generación de los primeros modelos de clasificación con valores de AUCs de 0,74 en el conjunto de descubrimiento y de 0,54 en el de replicación (382). Los resultados observados aún están lejos de ser aplicables tanto a nivel clínico como forense, pero son el primer paso de un largo camino que ya hemos comenzado a recorrer, camino aún por descubrir para el resto de las drogas duras. Si pensamos en otras drogas como la cocaína, heroína u opiáceos el panorama es muy diferente, la prevalencia de consumo es menor (1,4%, 0,0% y 4,0%, respectivamente) y los estudios que evalúan el consumo de estas sustancias son mucho más escasos. Para las drogas enumeradas se han identificado patrones de metilación diferentes en consumidores y no consumidores en

muestras de sangre (387–391) y, al igual que en caso anterior, se han evaluado ciertas características de interés para el desarrollo de modelos de clasificación. En el caso del consumo de cocaína y heroína se han identificado posiciones asociadas al tiempo que una persona lleva inyectándose y la intensidad de dichas inyecciones (392). Por otro lado, posiciones en el gen *OPRM1* presentan diferentes grados de metilación en base a la ancestralidad de los individuos en consumidores de opiáceos (393).

La identificación de consumidores de alguna de estas drogas está lejos de ser una realidad forense, pero su factibilidad ha sido demostrada. Nos encontramos ante una fase de descubrimiento, donde se requiere un mayor número de estudios, con un número significativo de consumidores y no consumidores, que permitan llevar a cabo una selección de marcadores específicos del consumo de cada una de las sustancias de interés en diversos tejidos. Este proceso puede ser largo y complicado, el reclutamiento de participantes, a diferencia de en otras aplicaciones, puede llegar a ser una gran limitante, pero la generación de un buen conjunto de datos es un requisito *sine qua non* para la creación de buenos modelos de clasificación.

d) Ejercicio físico

Combatir el sedentarismo ha sido un objetivo de salud importante en los últimos años, con el auge de las nuevas tecnologías y la creación de un mayor número de trabajos estacionarios pasamos una mayor cantidad de horas al día en posiciones estáticas. El impacto del ejercicio fisco en nuestra salud ha sido ampliamente estudiado a nivel clínico, demostrando los efectos positivos del deporte en la prevención o tratamiento de enfermedades (394–396), observándose tasas de mortalidad de entre 3 a 5 veces inferiores en individuos de la misma edad que no realizan actividad física (397). La asociación ha quedado ampliamente demostrada, pero ¿qué pasa con los mecanismos que hacen realidad esta mejora de salud? En la búsqueda de la resolución a esta incógnita surge el estudio de un biomarcador, influenciado por el ambiente, que modifica la expresión génica sin alterar la secuencia del ADN, la metilación del ADN. Los primeros estudios realizados demostraron un modificación global o gen-específica de la metilación del ADN en músculo esquelético producida por la actividad física (398). Gracias a este acercamiento se ha abierto la puerta a diversos estudios que evalúan los efectos beneficiosos de la actividad física en diversas enfermedades (399–402) o durante la etapa de recuperación mejorando la respuesta a los tratamientos, como por ejemplo en el caso de cáncer (403).

El estudio de este estilo de vida, a parte del ya descrito, presenta aplicaciones interesantes extrapolables al ámbito forense. La predicción de la actividad física en el campo forense se sitúa como una posible nueva *DNA intelligence tool*, aportando información que podría ser de interés para reducir el número de sospechosos. La pregunta lógica que uno puede hacerse tras esta afirmación es ¿en qué medida puede proporcionar información que una persona tenga un grado alto/bajo/nulo de actividad física? La figura de una persona suele estar asociada a su índice de masa corporal (BMI), por tanto, el estudio de la relación de la actividad física y el BMI podrían proporcionar información sobre el tipo de cuerpo del donante de una muestra. Esto es algo teórico, y es altamente probable que otros factores, como la dieta, deban ser estudiados en

paralelo para identificar asociaciones en los patrones de metilación específicas del ejercicio físico. El estudio de este índice ha mostrado una interesante relación con la edad epigenética, observándose en gemelos que 10 unidades de BMI corresponden con 1-3 años de edad biológica en sangre, hígado y tejido adiposo (404–406). Una correlación similar se observa al estudiar actividad física y edad epigenética, demostrándose que un alto nivel de ejercicio está asociado con un epigenoma más joven, mientras que la inactividad se asocia con una aceleración del envejecimiento epigenético (407). Aunque la aplicación es muy interesante, su literatura, por el momento, es demasiado heterogénea y compleja, limitando la consistencia de los resultados observados a diversos factores como: número pequeño de muestras, poblaciones heterogéneas, diversas clasificaciones de grado de actividad física y diferentes tejidos (408). Por tanto, la aplicación más cercana y realista de este conocimiento, a nivel forense, pasa por la evaluación de la influencia de este estilo de vida en los modelos de predicción que puedan verse influenciados, como, por ejemplo, la predicción de la edad.

e) Dieta

La actividad física está altamente relacionada con otro estilo de vida de interés, la alimentación. Esto se refleja en la intrínseca relación entre la obtención de buenos resultados a nivel físico y una dieta equilibrada y saludable. Los efectos positivos en la salud de una correcta alimentación han sido ampliamente estudiados y demostrados, siendo un hábito de gran interés y preocupación a nivel mundial debido al aumento del consumo de alimentos hipercalóricos. Por tanto, los estudios sobre el tipo de alimentación recomendada han aumentado paulatinamente, más desde el descubrimiento de la metilación del ADN y sus primeras asociaciones con la nutrición (409). El estudio del efecto de la dieta sobre esta marca epigenética pasa por identificar los efectos producidos por los componentes de los alimentos en las rutas de expresión o, en este caso, en las cadenas de procesamiento de la metilación/desmetilación del ADN. Entre los elementos objetivo de estudio se encuentra: los folatos y betaínas, cofactores en las rutas de metilación; el alfa cetoglutarato, vitamina C y vitamina A, cofactores de la desmetilasa TET; y cúrcuma, ácido rosmarínico, quertercina y luteolina, moduladores de las enzimas ADN metiltransferasas (409–411). Teniendo esto en cuenta se ha evaluado cómo la modificación de la dieta, centrándose en el consumo de estos componentes, puede alterar los patrones epigenéticos (412), observándose incluso cambios en genes relacionados con el metabolismo, crecimiento y regulación del apetito de recién nacidos como consecuencia de la alimentación de la madre antes y durante el embarazo (413). Estas modificaciones en patrones epigenéticos han sido asociadas a enfermedades como el cáncer, alterándose genes asociados a la regulación de la proliferación celular, apoptosis e inflamación (143). En el ámbito forense, la aportación de información relativa a la dieta está lejos de ser una realidad, y más por sí sola. La capacidad de discriminación que nos aportaría el conocimiento del tipo de dieta consumida por el donante de un vestigio biológico es limitada y muy dependiente del modelo establecido y los alimentos que tiene en cuenta. Pero esto no hace que el estudio de la alimentación sea menos interesante, su análisis en conjunto con el ejercicio físico y el BMI podrían aportar información sobre el aspecto corporal de un individuo, siendo una información mucho más relevante para reducir el número de sospechosos.

f) Enfermedades y modelos de predicción forenses

A lo largo de los distintos puntos tratados en esta tesis hemos observado, de forma superficial, la importante relación entre la metilación del ADN y el ambiente, siendo uno de sus resultados la predisposición a enfermedades. A nivel forense, el estudio de enfermedades está muy limitado a nivel judicial, considerándose que dicha información proporciona datos genéticos distintos de los meramente identificativos. Esta idea nos acompañó durante años limitando el estudio del ADN en este ámbito a las regiones no codificantes, entendiéndose el ADN codificante como poseedor de información sensible y externa a los intereses de las investigaciones. Según el Ministerio de Justicia de España esta percepción comienza a cambiar cuando se descubre que las regiones no codificantes son mucho más de lo que inicialmente se pensaba, conteniendo información genética distinta de la meramente identificativa (414). A esto debe sumársele el potencial, ya comentado, del ADN codificante para la identificación de rasgos fenotípicos o biogeográficos, de gran utilidad a nivel forense, que ha permitido el estudio de estas regiones del ADN con objetivos específicos. Ante esta realidad legislativa, el estudio de enfermedades con fines forenses queda fuera del alcance legal. Aun así, el análisis del efecto de las enfermedades sobre los patrones de metilación podría proporcionar información relevante sobre los efectos de estas condiciones en los modelos de predicción desarrollados.

El impacto de las enfermedades sobre el epigenoma ha sido estudiado con el objetivo de aportar contexto y responder incógnitas que el código genético, por sí solo, no puede resolver (415). La relación de la metilación del ADN y las enfermedades está presente en las aplicaciones forenses asociadas a este biomarcador, por tanto, es necesario, siempre que sea posible, evaluar sus efectos sobre los marcadores o modelos de clasificación o predicción desarrollados a fin de probar su robustez ante muestras desconocidas. Algunos estudios han realizado esta evaluación en sus modelos, pero debemos tener en cuenta que este hecho, aunque aporta información relevante, es específico de las posiciones a las que es aplicado y no puede ser extrapolado a otros marcadores. En relación con la identificación de tejidos, Forat et al. (206) evaluaron los efectos de diferentes tumores comunes sobre los marcadores seleccionados en su modelos de clasificación, observando diferencias destacables para algunos marcadores específicos de tejido en relación con alguno de los tumores analizados. En el contexto de la predicción de la edad se ha observado que marcadores correlacionados con la misma y empleados en modelos de predicción se ven alterados ante la presencia de enfermedades cardiovasculares, neurodegenerativas (289) y diabetes de tipo II (416), pero no se ha profundizado en su efecto sobre los modelos de predicción. Este evaluación directa sí que se ha llevado a cabo en otros estudios, mostrando escenarios diferentes en los que las enfermedades evaluadas parecen no influir en la predicción de la edad (262) y en los que influyen directamente incrementando los errores observados (273). En este último caso, es importante destacar que el modelo de sangre evaluado (± 3,8 años) presenta errores de ± 12,74 años en individuos que presentan enfermedades relacionadas con la sangre. Por otro lado, también se han identificado asociaciones entre marcadores relacionados con el consumo de tabaco y alcohol con enfermedades cardiovasculares y, en concreto, el consumo de tabaco con cáncer de pulmón (340). Con estos acercamientos se destaca la importancia de estos análisis, pero es entendible

que en muchas ocasiones estas evaluaciones no puedan llevarse a cabo por falta de información o escasez de muestras. A fin de solventar esta limitación sería recomendable que se realizasen evaluaciones más generales como la aproximación llevada a cabo por Silva et al. (417). En dicho trabajo evaluaron el efecto de cinco desordenes de la salud sobre 137 posiciones CpGs tejido específicas recopiladas de la literatura. De las posiciones evaluadas 45 mostraron diferencias significativas entre casos y controles, teniendo en cuenta los resultados obtenidos para los cinco conjuntos de datos empleados.

Enfrentarse a una muestra desconocida puede implicar muchas asunciones que pueden tener efectos muy significativos en los resultados obtenidos, como hemos visto en este apartado. A fin de reforzar la robustez de los modelos desarrollados en genética forense, sería recomendable la realización de estudios que evaluasen los efectos de distintas enfermedades sobre los marcadores empleados en las aplicaciones forenses del estudio de la metilación del ADN.

OBJETIVOS

2. OBJETIVOS

Objetivo general:

El objetivo general de esta tesis doctoral ha sido el desarrollo de modelos de predicción a partir de marcadores epigenéticos (metilación del ADN) para la estimación de la edad cronológica individual, identificación de tejidos e inferencia de estilos de vida con fines forenses.

Objetivos específicos:

Para el desarrollo de este objetivo general, se han planteado los siguientes objetivos específicos:

i. Identificación y selección de marcadores epigenéticos correlacionados con la edad en diversos tejidos biológicos tales como sangre, mucosa oral, saliva y cartílago.

ii. Identificación y selección de marcadores epigenéticos específicos de tejido para mucosa oral, saliva, sangre, hueso y cartílago.

iii. Identificación y selección de marcadores epigenéticos relacionados con el consumo de tabaco y alcohol.

iv. Diseño y optimización de un ensayo experimental para el análisis de los niveles de metilación en un conjunto de marcadores epigenéticos mediante minisecuenciación.

v. Validación de los marcadores epigenéticos seleccionados para estimación de la edad, identificación de tejido e inferencia de consumo de tabaco y alcohol.

vi. Desarrollo de modelos de predicción de la edad en varios tejidos forenses (sangre, mucosa oral, saliva y cartílago) mediante el análisis estadístico y bioinformático de datos de metilación de ADN.

vii. Desarrollo de modelos de identificación de tejido (mucosa oral, saliva, sangre, hueso y cartílago) mediante el análisis estadístico y bioinformático de datos de metilación de ADN.

viii. Desarrollo de modelos de inferencia de estilos de vida (consumo de tabaco y alcohol) mediante el análisis estadístico y bioinformático de datos de metilación de ADN.

METODOLOGÍA

3. METODOLOGÍA

Durante el desarrollo de esta tesis doctoral, a fin de llevar a cabo los estudios que en ella se plasman, se han empleado diversas metodologías tanto técnicas como estadísticas para analizar la metilación del ADN. El uso de muestras y su análisis ha sido garantizado mediante la obtención de la respectiva y necesaria aceptación de los comités de ética competentes (CAEI: 2013/543 y KBET/122.6120.86.2017). A continuación, se describen los métodos empleados en los diversos pasos del análisis a fin de proporcionar contexto a los resultados posteriormente presentados.

3.1. Metodología técnica

3.1.1. Extracción de ADN

Para el análisis de las muestras recogidas se han empleado dos métodos de extracción diferentes, un método de extracción líquido-líquido empleando fenol/cloroformo y otro en fase sólida empleado columnas de sílice. Tras la lisis de las muestras, la extracción fenol/cloroformo se basa en la inmiscibilidad de dichos compuestos y el agua, generándose una fase acuosa y otra orgánica entre las que se separan las diferentes moléculas presentes en la muestra. La generación de estas dos fases de distintas densidades y composición llevan a la separación de los compuestos hidrofóbicos, contenidos en la fase orgánica; las proteínas, en la interfase; y los ácidos nucleicos, en la fase acuosa. La recolección de esta fase acuosa y la repetición de esta separación permite aumentar la pureza del ADN extraído. Esta técnica proporciona una extracción con bajo coste, pero es un proceso relativamente largo y utiliza reactivos peligrosos que requieren de unas condiciones de trabajo específicas. Por otro lado, el método de extracción basado en sílice, empleando en este caso el *kit Sherlock AX* de A&A Biotechnology, se centra en la unión selectiva, bajo determinadas condiciones, de los ácidos nucleicos a una membrana o fase sólida. Una vez unida se realizan lavados, lo que permite eliminar de forma efectiva los inhibidores de la PCR presentes en la muestra obteniéndose un extracto purificado, y posteriormente, se llevará a cabo la elución del ADN. Este proceso de extracción permite obtener las muestras extraídas en poco tiempo y en pocos pasos, facilitando su implementación en ámbitos como el campo forense.

3.1.2. Cuantificación de ADN

Una vez finalizada la extracción es necesario conocer la cantidad de ADN extraído, a fin de definir la cantidad de muestra empleada en los pasos posteriores. Para ello se han empleado dos metodologías diferentes según el trabajo: cuantificación mediante espectrofotometría o fluorescencia. La primera de ellas, empleando el espectrofotómetro *NanoDrop 8000 UV-Vis*, se basa en el principio de absorbancia, definida como la medida de la atenuación de una radiación, del espectro ultravioleta visible, al atravesar una sustancia. La concentración de ADN se calcula, en este caso, a partir de los valores de absorbancia medidos a la longitud de onda deseada. Esta metodología permite estudiar la pureza de las muestras, pero estas impurezas pueden no ser correctamente detectadas llevando a una sobreestimación de la concentración.

Por otro lado, los fluorímetros, empleando en este caso los *kits Qubit dsDNA High Sensitivity* (HS) y *dsDNA Broad Range* (BR) de ThermoFisher, permiten la cuantificación de ADN mediante la detección de fluorocromos que se hibridan de forma específica al ADN mono o bicatenario. Este proceso tiene una mayor sensibilidad que el anteriormente descrito y su gran rapidez facilita su implementación y aplicabilidad. Aun así, debe tenerse en cuenta que no es específico de ADN humano y la contaminación con otras especies es indistinguible en este paso, pudiendo condicionar la cantidad detectada y la muestra usada en procesos posteriores.

3.1.3. Conversión con bisulfito sódico

La conversión con bisulfito sódico, como hemos visto en la introducción, es un paso fundamental en la mayoría de las metodologías empleadas para el análisis de la metilación del ADN. El uso de *kits* comerciales, como los empleados en el desarrollo de esta tesis doctoral (*MethyEdge® Bisulfite Conversion System* de Promega, *EZ DNA Methylation kit* y *EZ DNA Methylation-Direct kit* de Zymo Research), han permitido la conversión de las muestras analizadas partiendo de una cantidad de 100 ng, 300 ng y 500 ng, según el trabajo. Si bien es cierto que este proceso ha sido optimizado, la conversión con bisulfito es un tratamiento agresivo que puede producir la degradación del ADN como consecuencia de los largos tiempos de incubación, las altas temperaturas y una alta concentración de bisulfito. Teniendo esto en cuenta, este paso se convierte en un punto crítico del análisis donde la degradación de la muestra o una conversión incompleta pueden tener grandes efectos en los valores detectados.

3.1.4. Análisis de la metilación del ADN

Como ya se ha comentado existen diversos métodos que nos permiten analizar la metilación del ADN. En el transcurso de esta tesis doctoral se han desarrollado modelos en base a datos obtenidos con tres de las técnicas más utilizadas en el campo forense, minisecuenciación, EpiTYPER y MPS. Los valores de metilación generados por dichas plataformas fueron obtenidos de formas diferentes: i) produciendo los datos en el laboratorio durante el desarrollo de la presente tesis doctoral; ii) empleando datos publicados previamente generados en el laboratorio; iii) empleando datos obtenidos mediante la colaboración con un laboratorio externo; iv) solicitando acceso a la base de datos *UK AIRWAVE study*. Teniendo en cuenta este contexto, a continuación, se describen las metodologías empleadas para la generación de dichos datos.

El proceso de minisecuenciación se llevó a cabo con el *kit SNaPshot™ multiplex*, que permite realizar una extensión de un solo nucleótido marcado con una molécula fluorescente. La detección de esta fluorescencia y, por consiguiente, del grado de metilación del marcador analizado se llevó a cabo empleando un secuenciador de electroforesis capilar *ABI3130xl Genetic Analyzer* de Applied Biosystems. Por otro lado, la detección y cuantificación de la metilación del ADN de las muestras analizadas con EpiTYPER se llevó a cabo empleando la espectrometría de masas *MassARRAY®*. Por último, se emplearon datos generados con tecnologías de secuenciación masiva en paralelo de Illumina, MiSeq® y el array *Infinium MethylationEPIC BeadChip*. Para la obtención de los datos con el sistema MiSeq® se

construyeron las librerías empleado los *kits KAPA Hyper Prep Kit* y *KAPA Unique-Dual Indexed Adapters* de Roche, realizándose la secuenciación con el *kit MiSeq® Reagent Kit v3* de Verogen. Por último, los datos solicitados en la base de datos *UK AIRWAVE study* fueron generados con el *chip* de Illumina array *Infinium MethylationEPIC BeadChip*, proporcionando información de 853.307 CpGs.

3.2. Metodología estadística

Los análisis estadísticos llevados a cabo para evaluar las tendencias y comportamiento de los patrones de metilación, así como para la construcción de los modelos de predicción presentados en esta tesis doctoral, se realizaron empleando el *software* libre R. Este sistema es un lenguaje y ambiente diseñado para la computación y representación gráfica estadística que proporciona una amplia variedad de técnicas. A su vez, ofrece a los usuarios la posibilidad de construcción de sus propias funciones y funcionalidades lo que proporciona una mayor diversidad tanto en el tratamiento como en el análisis de los datos. Este sistema se ha aprovechado empleando tanto las funciones básicas como paquetes específicos para la interpretación, análisis y representación de los valores de metilación estudiados. Al ser un *software* en constante desarrollo, a lo largo del tiempo, se ha hecho uso de diversas versiones de R, empleando desde la v3.4.2 hasta la v4.2.2 (418).

3.2.1. Patrones de metilación y selección de marcadores

La construcción de modelos de predicción en base a metilación del ADN es dependiente de un buen estudio de los patrones de dicho biomarcador y de una correcta selección de los marcadores más informativos para cada condición evaluada. Por tanto, debe decidirse correctamente que estadísticos emplear a la hora de evaluar la correlación, el poder de clasificación o definir qué modelos son los más recomendables para el tipo aplicación. Para llevar a cabo estas comprobaciones es necesario evaluar la distribución de los datos o emplear estadísticos no paramétricos, en los que no se define una distribución *a priori* y no se asume que los datos deban ajustarse a una distribución concreta. La normalidad de las posiciones CpG analizadas se evaluó empleando la prueba de *Shaphiro-Wilk* en los residuos de los modelos lineales generados con las posiciones bajo estudio. Al no presentar una distribución normal, o si ésta no fue evaluada, se aplicaron métodos estadísticos no paramétricos, como la correlación de *Spearman* o la prueba de *Wilcoxon-Mann-Whitney*. El coeficiente de correlación de *Spearman* se empleó a fin de evaluar la correlación de la metilación y la edad. Los valores obtenidos con dicho coeficiente se encuentran entre -1 y +1, definiendo una asociación negativa o positiva entre las variables interrogadas, respectivamente, y una no correlación cuando los valores son cercanos o iguales a 0. Por otro lado, para evaluar la independencia de alguna variable, como por ejemplo el efecto del sexo biológico, se empleó la prueba de *Wilcoxon-Mann-Whitney*, prueba no paramétrica de la t de *Student*. Por último, la replicabilidad y la dispersión de los marcadores seleccionados se evaluó calculando la desviación típica entre réplicas o entre las muestras comprendidas en las categorías definidas en el análisis. Una desviación baja hace referencia a que la mayor parte de los datos presentan valores cercanos a

la media de la distribución, mientras que en el caso contrario los valores se extienden en un rango más amplio.

3.2.2. Modelos de predicción

Durante el transcurso de esta tesis doctoral se han empleado diferentes metodologías estadísticas para el desarrollo de diversos modelos de predicción. Los modelos de predicción de la edad, donde se evalúan dos variables continuas (edad y patrones de metilación), se desarrollaron empleando regresión cuantil (QR), regresión cuantil aplicada a redes neuronales (QRNN) y regresión cuantil aplicada a vectores de soporte (QRSVM). La regresión cuantil es una técnica no paramétrica que modela la relación entre dos variables estimando la mediana y otros cuantiles de la variable respuesta. Esta metodología no hace asunciones sobre la distribución de la variable objeto y las estimaciones proporcionadas, al estar basadas en la mediana, son más robustas frente a valores atípicos. Dicha técnica es aplicable a máquinas de aprendizaje como redes neuronales y vectores de soporte, modelos de inteligencia artificial que, empleando algoritmos, identifican patrones y elaboran predicciones a partir de ellos. Para los modelos de regresión cuantil generados se definen cuantiles que dividen la distribución en grupos. En nuestro caso se emplean los cuantiles q10, q50 (mediana) y q90 a fin de evaluar el 80% de la distribución generada por el modelo, definiéndose por tanto estos intervalos de confianza para cada observación. El cuantil 10 hace referencia al valor bajo el cual se encuentran el 10% de los valores más bajos y el cuantil 90 hace referencia al valor sobre el que se encuentran el 10% de los valores más altos. Una vez generados los modelos, deben definirse los parámetros estadísticos que van a evaluarse para interpretar los resultados obtenidos. Con el objetivo de evaluar la eficacia de los modelos generados se evalúan el error medio absoluto (MAE_{mean}), el error mediano absoluto (MAE_{median}), la raíz del error cuadrático medio (RMSE) y el porcentaje de predicciones correctas dentro de los intervalos de predicción (%CP±PI). Estas métricas permiten evaluar el rendimiento de los modelos generados ofreciendo información a partir del análisis de la variable real analizada, en este caso la edad cronológica, y las predicciones generadas por el modelo, la edad predicha. Con esto en mente, el MAE_{mean} y MAE_{median} proporcionan la media y la mediana, respectivamente, de las diferencias entre la variable real y la predicha y, por otro lado, el RMSE representa la media cuadrática de dichas diferencias. Por último, el porcentaje de clasificaciones correctas dentro del intervalo de predicción hace referencia, para el conjunto de datos analizado, al número de valores predichos que se encuentran dentro del rango intercuantil (q10 y q90).

Por otro lado, para clasificar una variable categórica (variable con un número limitado de categorías), como por ejemplo tipo de tejido o estilos de vida, en función de variables predictoras (metilación del ADN) se empleó la regresión logística binomial, cuando la variable categórica es binaria, o multinomial, cuando la variable categórica está compuesta por más de dos categorías. A fin de evaluar los modelos logísticos desarrollados se evalúa el área bajo la curva (AUC), la sensibilidad, especificidad y porcentaje de clasificaciones correctas del modelo. El AUC es una métrica que permite evaluar cómo de bien se clasifican las observaciones positivas y negativas en modelos de regresión logística, presentando los mejores

modelos valores de AUC cercanos a uno. Por otro lado, la sensibilidad y especificidad evalúan la capacidad del modelo para predecir correctamente las observaciones positivas y las negativas, respectivamente. Por tanto, la sensibilidad hará referencia al porcentaje de positivos verdaderos y la especificidad al porcentaje de negativos verdaderos. A modo de ejemplo, teniendo en cuenta la variable fumar y las categorías fumador y no fumador; la sensibilidad nos proporciona el número de fumadores clasificados como fumadores por el modelo y la especificidad el número de no fumadores clasificados como no fumadores. Por otro lado, el porcentaje de clasificaciones correctas es otra métrica que nos permite evaluar estas clasificaciones, haciendo referencia al número de individuos correctamente clasificados en las categorías analizadas. Por último, para la visualización gráfica de estos modelos se ha empleado el análisis de componentes principales (PCA). Se genera una representación gráfica que representa una combinación lineal de las variables que explican la mayor cantidad de varianza (medida de dispersión) para las categorías evaluadas. Las componentes se pueden entender como dichas variables y la dirección que maximiza la varianza de los datos representados. Por tanto, a mayor varianza, mayor separación entre los grupos estudiados, lo que se traduce en una mejor clasificación.

Los modelos desarrollados fueron evaluados con conjuntos de datos de testeo en los casos en los que había muestras disponibles. De modo adicional, para todos los modelos se realizó una validación empleando el propio conjunto de entrenamiento. Teniendo en cuenta el número de muestras que componen el modelo se emplearon dos tipos de validaciones: validación cruzada y validación uno-fuera. La validación cruzada consiste en dividir los datos en un número definido de porciones iguales (en nuestro caso 10) y enfrentar una de ellas al resto del conjunto de datos. Por tanto, teniendo en cuenta las porciones empleadas en nuestros análisis, se genera el modelo con nueve décimos de las muestras del conjunto de entrenamiento y la décima parte sobrante se emplea como grupo de validación. Este proceso se repite con cada una de las porciones generadas y se calcula una media de los estadísticos evaluados. Por otro lado, cuando el conjunto de datos empleado en el desarrollo del modelo no es lo suficientemente grande para poder generar estas porciones, se emplea la validación uno-fuera. Dicha validación consiste en separar una muestra del conjunto de entrenamiento y emplearla como muestra de validación. Este proceso se repite un número de veces igual al número de muestras en el conjunto evaluado. La media de los resultados obtenidos teniendo en cuenta todas las validaciones se representa para los estadísticos analizados.

RESULTADOS

4. RESULTADOS

Artículo 1: A common epigenetic clock from childhood to old age

A. Freire-Aradas, L. Girón Santamaría, A. Mosquera-Miguel, A. Ambroa-Conde, C. Phillips, M. Casares de Cal, A. Gómez-Tato, J. Álvarez-Dios, E. Pospiech, A. Aliferi, D. Syndercombe Court, W. Branicki, M. V. Lareu.

Forensic Science International: Genetics, volumen 60, página 102743, 2022, PMID: 35777225, DOI: 10.1016/j.fsigen.2022.102743

Forensic Science International: Genetics 60 (2022) 102743

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

A common epigenetic clock from childhood to old age

A. Freire-Aradas [a,*], L. Girón-Santamaría [a], A. Mosquera-Miguel [a], A. Ambroa-Conde [a], C. Phillips [a], M. Casares de Cal [b], A. Gómez-Tato [b], J. Álvarez-Dios [b], E. Pospiech [c], A. Aliferi [d], D. Syndercombe Court [d], W. Branicki [e], M.V. Lareu [a]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] Faculty of Mathematics, University of Santiago de Compostela, Spain
[c] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[d] King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom
[e] Laboratory of Anthropology, Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland

ARTICLE INFO

Keywords:
DNA methylation
Forensic age estimation
Quantile regression
Machine learning
EpiTYPER®

ABSTRACT

Forensic age estimation is a DNA intelligence tool that forms an important part of Forensic DNA Phenotyping. Criminal cases with no suspects or with unsuccessful matches in searches on DNA databases; human identification analyses in mass disasters; anthropological studies or legal disputes; all benefit from age estimation to gain investigative leads. Several age prediction models have been developed to date based on DNA methylation. Although different DNA methylation technologies as well as diverse statistical methods have been proposed, most of them are based on blood samples and mainly restricted to adult age ranges. In the current study, we present an extended age prediction model based on 895 evenly distributed Spanish DNA blood samples from 2 to 104 years old. DNA methylation levels were detected using Agena Bioscience EpiTYPER® technology for a total of seven CpG sites located at seven genomic regions: *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429 (GRCh38). The accuracy of the age prediction system was tested by comparing three statistical methods: quantile regression (QR), quantile regression neural network (QRNN) and quantile regression support vector machine (QRSVM). The most accurate predictions were obtained when using QRNN or QRSVM (mean absolute prediction error, MAE of ± 3.36 and ± 3.41, respectively). Validation of the models with an independent Spanish testing set (N = 152) provided similar accuracies for both methods (MAE: ± 3.32 and ± 3.45, respectively). The main advantage of using quantile regression statistical tools lies in obtaining age-dependent prediction intervals, fitting the error to the estimated age. An additional analysis of dimensionality reduction shows a direct correlation of increased error and a reduction of correct classifications as the training sample size is reduced. Results indicated that a minimum sample size of six samples per year-of-age covered by the training set is recommended to efficiently capture the most inter-individual variability..

1. Introduction

Epigenetics plays a key role in the control of gene expression [1]. Epigenetic signatures affecting this molecular process are reversible, act in cascade or network and affect DNA regulation without altering the underlying DNA sequence [2]. Four main categories of epigenetic marks have been described: chromatin remodeling [3], post-translational histone modifications [4], non-coding RNAs [5] and DNA methylation [6]; with the latter the most widely studied so far. DNA methylation is the addition of a methyl group in the 5′ carbon of those cytosine residues predominantly located in CpG dinucleotides, that generally contributes to gene silencing [7,8]. A plethora of genome-wide studies have shed light on the DNA methylation process during the last ten years, many of them targeting CpG sites correlated with individual age [9–17]. Gradual age-correlated hyper- and hypomethylation patterns have been observed in the human genome [18]. Based on these observed correlations, a new concept termed "epigenetic age" emerged. Epigenetic age refers either to chronological or biological age depending on the marker set used. Additionally, depending on the individual's lifestyle and/or presence of disease, chronological or biological age might match or

* Corresponding author.
E-mail address: ana.freire@usc.es (A. Freire-Aradas).

https://doi.org/10.1016/j.fsigen.2022.102743
Received 21 December 2021; Received in revised form 22 June 2022; Accepted 23 June 2022
Available online 25 June 2022

differ in scale. While chronological age has been proved to be useful in a forensic context [19], biological age might also be used to monitor the progress of a person with illness or undergoing treatment for a medical condition [20–22].

A universal epigenetic clock was proposed by Horvath in 2013 [13]. In spite of the advantages that such an age prediction model presented, covering multiple tissues in donors that ranged in age from newborns to centenarians and trained on more than 7000 control individuals, the analysis of more than 300 CpG sites hampered its application in platforms apart from Illumina HumanMethylation Beadchips [23]. To apply epigenetic clocks using alternative DNA methylation technologies, a substantial reduction of markers has been the strategy of choice for forensic applications [24].

In recent years, multiple age prediction models have been developed for forensic analysis using a reduced number of CpG sites. These epigenetic clocks were designed targeting multiple forensic tissues: blood [25,26], saliva [27,28], semen [29], teeth [30] and bones [31]; using a variety of technologies: Pyrosequencing [24,32], EpiTYPER [26,33], SNaPshot [27,28] or Massively Parallel Sequencing [34–37]; and applying different statistical models, including linear regression [25], quantile regression [38], support vector machine [39] or artificial neural networks [40]; as well as covering different age ranges: adults [41] and children [42]. Common to all of them is the use of a reduced number of markers, from 3 to 16 CpG sites.

Most age prediction models published to date mainly cover adult samples, with subjects below adult ages consistently underrepresented. Differences have been observed between children and adults in terms of epigenetic changes. DNA methylation patterns for some CpG sites, reveal a logarithmic dependence until adulthood that slows to a linear dependence later in life, as depicted by Horvath [13]. This increased variation of epigenetic states during the early stages of life could be explained by the rapid maturation of the immune system at this period [43]. Nevertheless, some CpG sites present a linear or quasi-linear pattern of gradual DNA methylation changes from childhood to very old age, which makes these the most suitable epigenetic biomarkers for establishing a common age prediction model that includes all age ranges, that statistically can be treated in a unified way.

To develop a common epigenetic clock covering the whole lifetime of a person, inter-individual variability should be also considered. Since epigenetics is the result of environmental interaction with genetics, individuals presenting similar chronological ages can be represented by multiple scenarios [44], including potential differences among populations [45]. Age prediction models reported so far have been trained using dozens to hundreds of volunteers, but no minimum sample size has been established to date.

In the present study, a common epigenetic clock for all human ages – from children to centenarians – was developed using seven CpG sites detected using EpiTYPER® technology. A total of 895 Spanish blood DNA samples ranging from 2 to 104 years old were trained exploring three statistical models. K-fold cross-validation was used for validation purposes, as well as an independent testing set composed of 152 Spanish individuals from 3 to 69 years old. Additionally, an optimal training set size was calculated assessing dimensionality reduction based on stepwise-reduced training sets from a total of 895 to 99 individuals.

2. Materials and methods

2.1. DNA samples and quantification

A total of 1047 blood DNA samples from Spanish donors (collected across multiple regions) ranging from 2 to 104 years old (mean age: 44.51 years, 477 males and 570 females, approximately 10 individuals per age) were used for the development of an extended age prediction model – DNA methylation levels for all samples were collected within the scope of previous projects. From these samples, 895 (~85%) were used for establishing the training set and the remaining 152 (~15%) for validation purposes. DNA samples were obtained from the Spanish National DNA Bank Carlos III, University of Salamanca and from the Bio-Bank IBSP-CV (PT13/0010/0064), integrated in the Spanish National Biobanks Network and in the Valencian Biobanking Network; and they were processed following standard operating procedures with the appropriate approval of the Ethical and Scientific Committees. Ethical approval for the present study was granted from the ethics committee of investigation in Galicia, Spain (CAEI: 2013/543). Additionally, two internal controls were included in all methylation analyses in order to confirm the reproducibility of results (blood DNA samples from a male and a female, 59 and 32 years old, respectively). All DNA samples were quantified by Qubit® dsDNA High Sensitivity (HS) Assay Kit (Thermo Fisher) and subsequently normalized to 10 ng/μL.

2.2. CpG target sites and Agena Bioscience EpiTYPER® DNA methylation analysis

The DNA methylation markers selected for this study were seven CpG sites from the genomic regions: *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429 (GRCh38), included in a previous age prediction model initially created for adult samples [26]. The Agena Bioscience EpiTYPER® system (San Diego, CA, USA) is a bisulfite-treatment-based method for detection and quantification of DNA methylation using MassARRAY® mass spectrometry [46]. The bisulfite conversion step is performed with the EZ DNA Methyl- ation™ Kit (Zymo Research), with input of 300 ng of genomic DNA, following manufacturer's guidelines, to produce 40 μL of converted DNA. That means that from the samples normalized to 10 ng/μL, 30 μL were used for bisulfite conversion. From the final 40 μL of converted DNA, 1 μL was used for subsequent EpiTYPER analyses. EpiTYPER DNA methylation data for the present study were obtained from two previous publications [26,42]. EpiTYPER detects methylation levels as CpG sets, comprising one or multiple CpG positions in the same cleavage fragment. Therefore, multiple CpGs in a set will be detected when they are closely positioned on the targeted fragment. Herein, we use the term CpG site whether one CpG, or a cluster of CpGs is detected in a single short DNA fragment.

2.3. Statistical analyses and establishment of online age prediction tools

All calculations were performed using R software v.3.4.2. Chronological age refers to the actual self-declared age of the individual. Correlations between DNA methylation levels and chronological age were calculated using the Spearman correlation (r_s). Inter-individual variability based on DNA methylation levels was assessed using standard deviation (threshold SD >0.05). To build an extended age prediction model, three statistical tools were explored using quantiles 0.1 and 0.9 (q10 and q90): quantile regression (QR), quantile regression neural network (QRNN) and quantile regression support vector machine (QRSVM); using the quantreg (rq function), qrnn (mcqrnn.fit function, 3 hidden layers) and liquidSVM (qtSVM function) R packages, respectively [47–49]. Validation of the prediction models was performed using k-fold cross-validation (k = 10) applying an R script developed in-house. The k-fold cross-validation randomly cleavages the input data (N = 895) into k fragments of similar sample size. Random cleavage of the input data was made using the cvTools R package [50]. Every k time that the model was assessed, a k cluster was retained as the test set with the remaining clusters used as the training set, maintaining proportions of 10% and 90% of the input data for test and training sets respectively, per run. The corresponding predictive accuracy was measured with the following performance metrics: mean absolute prediction error (MAE); root-mean-square error (RMSE) and percent of correct classifications within the prediction intervals (%CP±PI). Although when working with quantiles (QR, QRNN and QRSVM), the MAE can be represented by the median instead of the mean, the mean was used in the present study for comparative purposes with additional models, where the MAE is usually based on the mean. Correlation between epigenetic age and

chronological age was tested using R^2. Predicted versus chronological age was plotted using the ggplot2 R package [51]. To assess potential differences between statistical methods, a p-value < 0.05 was considered statistically significant. Potential sex differences were tested using the Wilcoxon Mann Whitney test. The final online age prediction tools (QRNN and QRSVM) developed in our study have been placed in the open-access *Snipper* forensic classification website and are freely available at: http://mathgene.usc.es/cgi-bin/snps/age_tools/processmethylation-blood_2–104.cgi.

3. Results

3.1. Association of seven CpG sites with chronological age using Agena Bioscience EpiTYPER®

Seven age-correlated CpG sites from the genomic regions: *ELOVL2* (CR_1_CpG_9: cg21572722), *ASPA* (CR_2_CpG_3: cg02228185), *PDE4C* (CR_4_CpG_27.28.29: none CpG_ID), *FHL2* (CR_12_1_CpG_3: cg06639320), *CCDC102B* (CR_13_CpG_2: cg19283806), *MIR29B2CHG* (CR_21_CpG_11: none CpG_ID) and chr16:85395429 (CR_23_CpG_3: cg07082267); were selected according to a previous age prediction model initially created for application to adult samples (CpG details in Table 3 of [26]). These epigenetic markers were analyzed in the present study using EpiTYPER® in a total of 895 individuals ranging from 2 to 104 years old. Reproducibility of results was confirmed using two internal controls that were included in all analyses. Fig. 1 represents the corresponding DNA methylation values compared with the chronological age. While *ELOVL2*, *PDE4C* and *FHL2* displayed hypermethylation with age; a decrease in the DNA methylation levels for *ASPA*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429 was observed. The strongest age-correlation was observed in *ELOVL2* (r_s: 0.97), followed by *MIR29B2CHG* (r_s: −0.95), *FHL2* (r_s: 0.94), *PDE4C* (r_s: 0.94), *CCDC102B* (r_s: −0.93), chr16:85395429 (r_s: −0.92) and lastly *ASPA* (r_s: −0.86). Inter-individual variability was similar at all ages in most markers (average SD <0.05), except *ASPA* (average SD: 0.068) and *MIR29B2CHG* (average SD: 0.063), which both gradually displayed inter-individual dispersion with increasing age, starting from the age of 40 and 30 years old, respectively.

3.2. Development and validation of a full coverage age prediction model based on EpiTYPER®: from children to the elderly

In view of the high age-correlation displayed by the markers assessed in Section 3.1, the seven CpG sites detected using *Agena Bioscience EpiTYPER®* were used to develop a common epigenetic clock from childhood to old age. Table 1 describes the performance metrics for the training set composed of 895 individuals ranging from 2 to 104 years old, comparing the QR, QRNN and QRSVM statistical models and based on k-fold cross-validation (average values representing the ten clusters). Cluster-specific cross-validations have been detailed in Supplementary Material S1. According to Table 1, QR provided a MAE: ± 3.75, RMSE: 5.23 and %CP±PI: 78.77%. Despite the accuracy of these results, the observed metrics were improved by applying additional models. Both QRNN and QRSVM displayed similar metrics: MAE: ± 3.36, ± 3.41; RMSE: 4.83, 4.78 and %CP±PI: 81.45%, 79.66%, respectively. Statistically significant differences between errors (p-value <0.05) were found between QR and QRNN, and between QR and QRSVM. However, when

Table 1
Performance metrics for the training, comprising 895 individuals, 2–104 years old and the testing set, comprising 152 individuals, 3–69 years; analyzing seven CpG sites (*ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429) detected with EpiTYPER®. The performance metrics for the training set are based on k-fold cross-validation (average values representing the ten clusters). MAE: mean absolute prediction error, RMSE: root-mean-square error, %CP±PI: percent of correct classifications within the prediction intervals, QR: quantile regression, QRNN: quantile regression neural network, QRSVM: quantile regression support vector machine.

Sample set	Model	MAE	RMSE	%CP±PI
Training	QR	± 3.75	5.23	78.77%
Training	QRNN	± 3.36	4.83	81.45%
Training	QRSVM	± 3.41	4.78	79.66%
Testing	QRNN	± 3.32	4.51	76.32%
Testing	QRSVM	± 3.45	4.75	77.63%

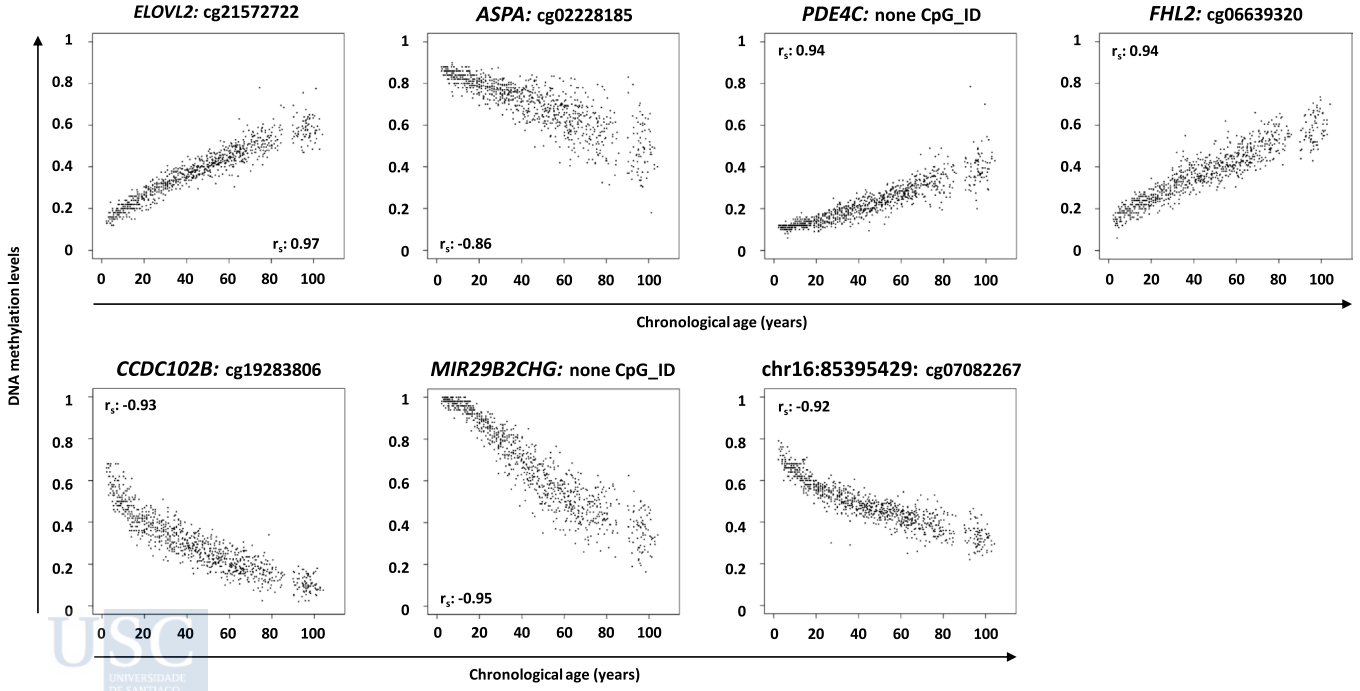


Fig. 1. Dispersion plots (DNA methylation levels compared with chronological age) for *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429 markers using DNA blood samples from 895 Spanish individuals (2–104 years old).

comparing errors obtained using QRNN and QRSVM, no statistically significant differences were found (p-value >0.05). Based on these results, subsequent analyses were constrained to QRNN and QRSVM. Fig. 2A and 2B show the epigenetic age versus the chronological age for QRNN and QRSVM models respectively. Continuous grey and black lines represent the perfect and fitted correlation, respectively; showing practical overlap between both lines in both models. A continuous correlation of the epigenetic age with the chronological age is evident from persons at early life stages to centenarians (R^2: 0.9669 for QRNN and 0.9679 for QRSVM). However, an increase in inter-individual variation is detectable with increasing age, observed as narrow prediction intervals (minimum and maximum predicted ages displayed by the discontinuous dark red lines) in children and young adults that gradually expand with age, representing the main reason why the consideration of both predicted age and the prediction interval improves accuracy of the reported results, avoiding a higher rate of misclassification. Sex was not detected as a confounder factor (p-value >0.05). The QRNN and QRSVM age prediction models are freely available from the open-access *Snipper* forensic classification website described in Material and Methods. Detailed information regarding the underlying data used for building the models implemented on the website can be found in Supplementary Material S2.

Further validation of the full coverage age prediction models was performed using a total of 152 independent blood samples ranging in age from 3 to 69 years old. Table 1 summarizes the corresponding performance metrics. Errors were similar to the corresponding training sets in QRNN and QRSVM models (MAE: ± 3.32, ± 3.45 and RMSE: 4.51, 4.75; respectively). The percentage of correct predictions, although similar, slightly diminished in scale in comparison to the corresponding training set %CP ± PI values (76.32% and 77.63%, respectively).

3.3. Dimensionality reduction of the training set

To develop the present epigenetic clocks, a total of 895 individuals were used for training the models. In order to capture a high variety of potential inter-individual differences, each year of age was represented by approximately 10 individuals. However, some accurate reported age prediction models have been developed using reduced training sets composed of an average one individual per year (i.e., using N ≈ 100). Aiming to establish an optimal sample size for training an age prediction model, and in order to get the minimum error covering the highest inter-individual variation, an assessment of stepwise-reduced training sets was performed considering an approximate total number of individuals per year of 10, 8, 6, 4, 2 and 1 (n_{train}) corresponding to a total number of samples in the training set of 895 (N = 441 males, N = 454 females), 707 (N = 334 males, N = 373 females), 552 (N = 253 males, N = 299 females), 379 (N = 171 males, N = 208 females), 195 (N = 89 males, N = 106 females) and 99 (N = 45 males, N = 54 females), respectively (N_{train}). Specific distributions per age and sex can be found for each analyzed training set in Supplementary Fig. S1-S6. Fig. 3 shows the MAE and correct classification rate within the prediction intervals using both QRNN and QRSVM models for the six explored N_{train} combinations. Table 2 summarizes the underlying performance metrics for the training set using the different N_{train} combinations based on k-fold cross-validation (average values representing the ten clusters). Cluster-specific cross-validations, as well as the corresponding learning curve plots can be found in Supplementary Material S3. According to Table 2, the MAE values from QRNN showed a gradual increase when decreasing the sample size from n_{train}= 10 to n_{train}= 1 (from ± 3.36 to ± 4.82). Although similar variations were found for the MAE from QRSVM, these were smaller in scale (from ± 3.41 to ± 4.07), providing a more stable error and an improved sample size independence. A gradual decrease in the QRNN correct classification rate (%CP ± PI) was observed accordingly when constraining the training set (from 81.45% to 60.67%). Similarly, when analyzing the same data with the QRSVM model, although a decrease was also detected, it was more stable across different N_{train} combinations (from 79.66% to 74%). Despite these variations between QRNN and QRSVM, no statistical differences were detected between models and across n_{train} numbers (p-value >0.05),

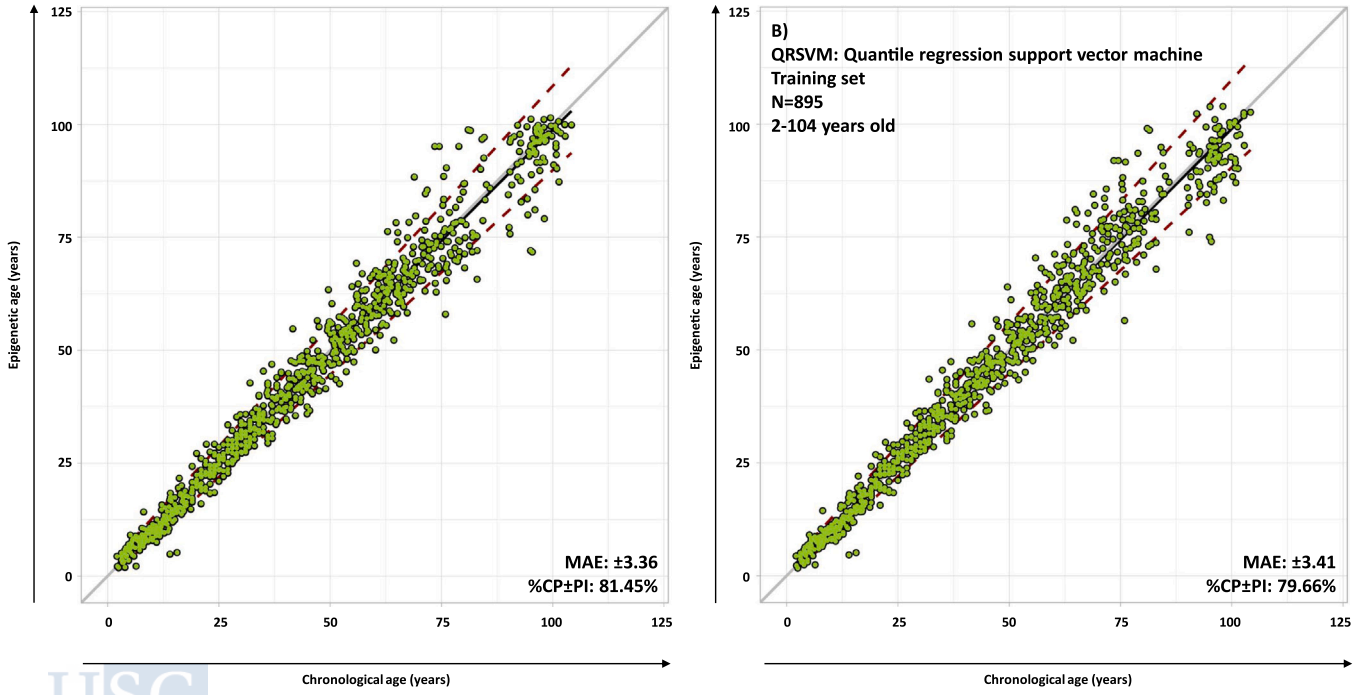


Fig. 2. Epigenetic age compared to chronological age for the training set of 895 Spanish individuals (2–104 years old) based on seven CpG sites in *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429 regions. The black diagonal line represents the 0.5 quantile regression line between epigenetic age and chronological age and the discontinuous (dark red) lines, the corresponding 0.1 and 0.9 quantile regression limits. The grey line is the diagonal line representing perfect correlation. The underlying statistical model was based on either QRNN (A) or QRSVM (B).

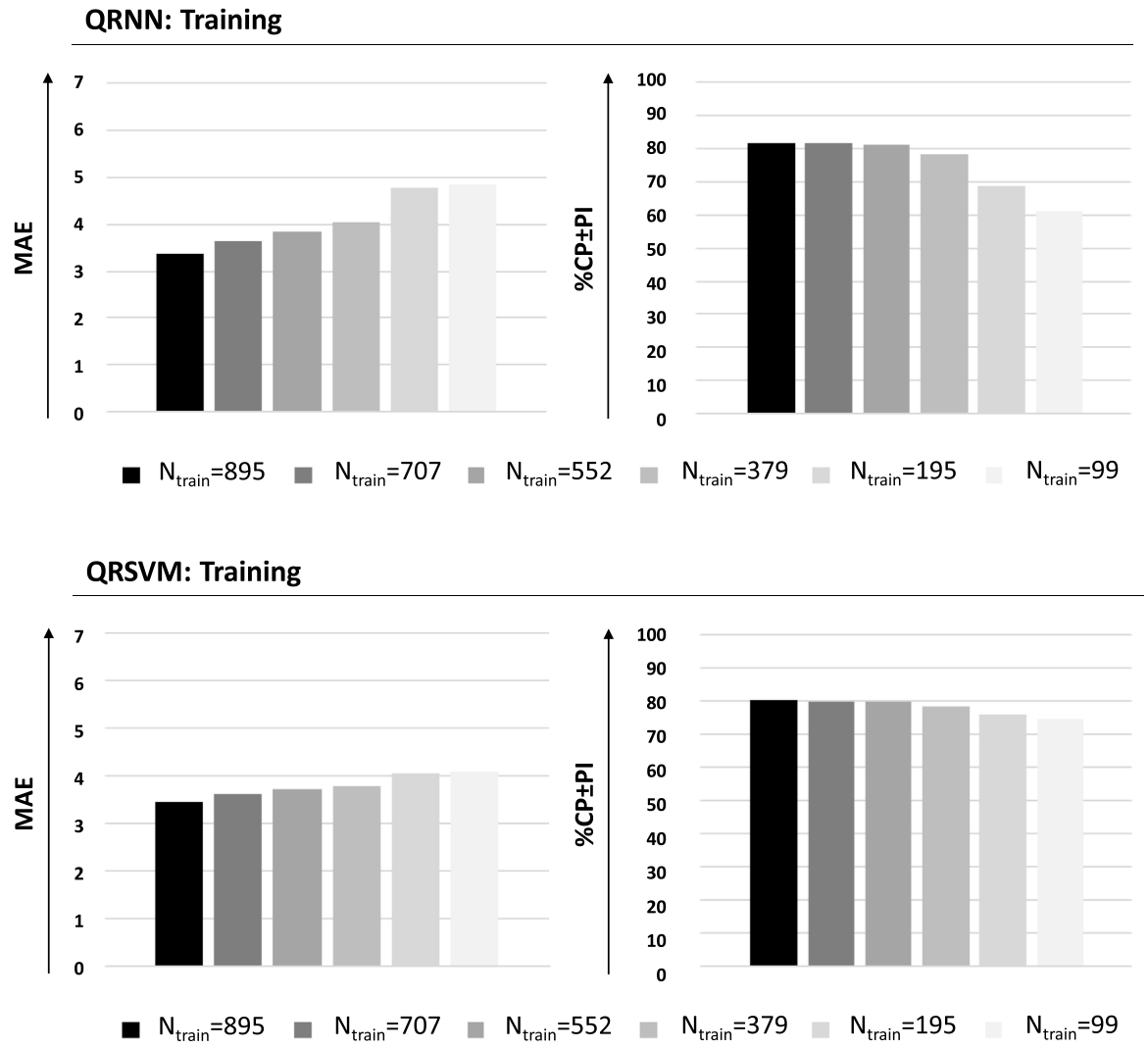


Fig. 3. Mean absolute prediction error (MAE) and percentage of correct classifications within the prediction intervals (%CP ± PI) using both QRNN and QRSVM models for the six N_{train} combinations evaluated.

Table 2
Stepwise assessment of six training sets, comprising samples of individuals in the age range 2–104 years old. The performance metrics for the training sets are based on k-fold cross-validation (average values representing the ten clusters). N_{train}: sample size of the training set, n_{train}: individuals per year at the training set, MAE: mean absolute prediction error, RMSE: root-mean-square error, %CP ± PI: percent of correct classifications within the prediction intervals, QRNN: quantile regression neural network, QRSVM: quantile regression support vector machine.

Model	N_{train}	n_{train}	MAE	RMSE	%CP ± PI
QRNN	895	10	± 3.36	4.83	81.45%
QRSVM	895	10	± 3.41	4.78	79.66%
QRNN	707	8	± 3.64	5.23	81.47%
QRSVM	707	8	± 3.58	4.99	79.48%
QRNN	552	6	± 3.82	5.44	80.96%
QRSVM	552	6	± 3.69	5.11	79.51%
QRNN	379	4	± 4.04	5.43	78.33%
QRSVM	379	4	± 3.76	4.94	77.8%
QRNN	195	2	± 4.77	6.69	68.76%
QRSVM	195	2	± 4.01	5.45	75.29%
QRNN	99	1	± 4.82	6.41	60.67%
QRSVM	99	1	± 4.07	5.28	74%

except for n_{train}= 1 (p-value: 0.0144).

To establish an optimal minimum sample size, it is important to consider that quantiles (q10 and q90) are applied in QRNN and QRSVM models by calculating the prediction intervals, which determine the percent of correct classifications (%CP ± PI). Since the interval between the quantiles applied is 80, about 80% of the samples should be correctly predicted to consider that the model is working properly. According to Fig. 3 and Table 2, a minimum sample size of n_{train}= 6, in this case corresponding to N_{train}= 552, would be the most appropriate sample size which can capture the most accurate age predictions (%CP ± PI≈80%) covering the broadest inter-individual variation. However, it is important to note that in terms of shrinking the training's sample size, QRSVM is less susceptible to lower numbers than QRNN (difference in MAE between N_{train}=895 and N_{train}=99: ± 1.46 and 0.66; difference in %CP ± PI between N_{train}=895 and N_{train}=99: 20.78% and 5.66%, for QRNN and QRSVM, respectively). The learning curve plots based on the MAE of each training set tested also show more stability when a minimum n_{train}= 6 is analyzed (Supplementary Material S3). Corresponding data for an independent testing set (N = 152) can be found in Supplementary Table S1. Although patterns are not exactly as displayed by Table 2, n_{train}= 6 continues to be the optimum sample size to be selected according to the previous criteria applied.

4. Discussion

Age estimation is a DNA intelligence tool aiming to provide additional information to the genetic profile at different scenarios, which can comprise: i) individual identification, ii) mass disaster screening, iii) forensic anthropology and iv) legal disputes about age. Subsequently, the development of epigenetic clocks is being implemented in forensic practice as a supplementary analysis for individual age prediction. In general terms, most of the forensic age prediction models to date have been based on adult samples [28,41], but minors and the variation in methylation patterns they show, must also be taken into account in the development of universally-applicable forensic analyses. Horvath's clock, although including both adults and minors, presents a different statistical treatment for both age ranges. While for ages below 20 years, a logarithmic transformation was applied, an untransformed linear model was used for ages above 20 years [13]. This logarithmic transformation is due to an exponential change on the DNA methylation levels at early stages of the individual's life [43]. Despite the high level of coverage of all ages used to develop Horvath's model, it is based on an impractically large number of markers – 353 CpG sites – representing a major drawback for forensic testing, due to the poor quality and/or quantity of DNA associated with most casework samples.

Minors have already been included in certain previous forensic epigenetic clocks, but these cover dispersed datapoints for age ranges under 18 years [24,25,34]. To improve this area of study, we developed a specific age prediction model for children and adolescents [42]. However, when a biological sample is found, no information is generally available to know if the donor is a minor or an adult. Therefore, the most useful epigenetic clock will be one unifying both age ranges into a single test model. A recent study from Wozniak et al., considered the whole range of ages from 1 to 75 years old (N = 112) to build a novel age prediction model [36], obtaining an MAE of ± 3.2 years for blood samples. In Wozniak's model, minors were represented from 0 to 18 years old with about one individual per year. Following this study, we aimed a step further by covering as much as possible the potential for inter-individual epigenetic variability. This was achieved by covering the fullest interval (2–104 years old) with approximately ten individuals per age (N = 895). By analyzing previously developed CpG sites in the seven genomic regions of *ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* and chr16:85395429, using EpiTYPER technology, we were able to retain a robust and efficient age prediction model [26], overlapping four of these loci with the recently released VISAGE Enhanced Tool for age prediction [36]. Additional CpG sites especially informative at childhood and adolescence had also been considered for analysis, such as *KCNAB3* [42]. However, aiming to build a single epigenetic clock covering all ages from childhood to the old age, a marker such as *KCNAB3* was discarded for analysis, because no linear correlation of DNA methylation levels with chronological age is maintained in this marker across all ages, exponentially increasing during childhood but presenting much more stable levels across adulthood (see Fig. 1 at [42]). This extreme lack of linearity between both age groups prevents the inclusion of this marker into a common epigenetic clock.

In addition to marker selection, although the age range and sample size of the training set were key factors in adapting our age prediction model, the underlying statistical model used also plays an important role. To date, application of linear regression [24,25,28,36,52] has been widely accepted. Since DNA methylation is quantitative in nature and gradually changes through the individual's lifetime, linear regression models fit well with age estimation based on this epigenetic signature. Quadratic regression models or power transformations have also been applied in cases where the change of the DNA methylation levels with chronological age demonstrates non-linear patterns [30,36]. Recently, novel statistical tools based on machine learning have been introduced [34,37,41,53]. Common to all these models is the fact that the error obtained is unique, independent of the age of the sample donor, and should be applied to whatever predicted age is achieved. Nevertheless, DNA methylation data consistently shows that young ages are better predicted than old ages and this has been observed in our dataset as well. Fig. 2 depicts the epigenetic age versus the chronological age for all the individuals from 2 to 104 years old. While the youngest subjects (under 20 years) have datapoints very closely positioned together, this pattern gradually changes until the oldest samples (over 80 years) that show the highest dispersion between datapoints. Inter-individual epigenetic variation is expected since epigenetics derives from an interaction between genetics and environment. Subsequently, when the longest period of time that two age-matched individuals have been exposed to different external factors applies, then major epigenetic differences will be encountered between them, in contrast to the earliest stages on life. In order to improve the accuracy of predictions, specific age-dependent errors could be applied if using statistical models based on quantile regression [26,38]. Inter-individual epigenetic variation could also occur among populations being affected by different environmental conditions at which the individuals are exposed to. In our study, we used a Spanish cohort in order to build and validate the age prediction model proposed. Further validation will be needed to demonstrate that our model can be used at different worldwide population groups.

In the present study, the QR, QRNN and QRSVM statistical prediction models have been tested for age estimation. The highest accuracy was obtained for QRNN in terms of MAE (± 3.36) and %CP ± PI (81.45%). Nevertheless, no statistical differences (p-value >0.05) were found between QRNN and QRSVM. Therefore, QRNN and QRSVM were both selected as the most accurate age prediction models and subsequent analyses were constraint to these two methods. Validation of the models with an independent set of samples (N = 152) produced similar results (MAE: ± 3.32, ± 3.45 and %CP ± PI: 76.32%, 77.63%; for QRNN and QRSVM, respectively). Nevertheless, since the testing set was restricted to 69 years old, further analysis of older samples should be required in order to validate these results in old age. Errors obtained in the present work were similar to previous models [25,28,34,41] (MAE: ± 3.9, ± 3.48, ± 4.1 and ± 3.24, respectively); nevertheless, these errors were fixed to whatever age was predicted. The main advantage of using age-dependent errors, such as those from quantile regression models used here, is to be able to narrow down the errors at early stages of life and to increase them at older ages, where inter-individual epigenetic variability plays an important role.

Finally, since the sample size of the training set is considered a key factor in developing an accurate age prediction model, the more samples are included, the more inter-individual epigenetic variation can be properly gauged, resulting in lower errors and a higher number of correct classifications. The stepwise dimensionality reduction we performed, taking into account that when quantiles are applied (q10 and q90), ~80% of samples should be correctly predicted, indicated N_{train}= 552 (6 individuals per year of age) gave an optimum balance between sample size and predictive accuracy. The sample size of six individuals per year provided for the training set, a MAE of ± 3.82 for QRNN and ± 3.69 for QRSVM. Correct classification rates were 80.96% for QRNN and 79.51% for QRSVM. At this analysis, it is important to note that, QRSVM showed to be less susceptible to shrinking of the training set than QRNN, therefore, in case of low number of samples, it could be used preferably. Similar metrics to the training were obtained for the corresponding testing set (MAE: ± 3.03, ± 3.56 and %CP ± PI: 84.87%, 78.29% for QRNN and QRSVM, respectively). However, patterns displayed by the testing set when tested under some of the stepwise-reduced training sets didn't follow exactly the same pattern as the corresponding training sets by themselves. This could be explained due to a reduced age range on the testing set (3–69 years old) in comparison to the training sets (2–104 years old). In summary, to cover a maximum level of inter-individual variability, six individuals per year of age is recommended for the development of future epigenetic clocks which aim to cover the complete range of human ages.

As a final remark, it should be taken into account that the underlying data used for building the age prediction models developed under this

study have been generated using EpiTYPER technology, a system that uses high quantities of genomic DNA (300 ng). In order to directly apply these models to forensic specimens usually presenting low quality and/or quantity of DNA, a step further will be to implement these age predictors on forensic technologies such as SNaPshot or Massively Parallel Sequencing, systems able to handle minor amounts of genomic DNA for methylation analyses of forensic casework.

Acknowledgements

AFA was supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010). The National DNA Bank Carlos III is supported by ISCIII, Ministry of Science and Innovation, Spain (PT13/0001/0037, PT13/0010/0067): The Murcia Twin Registry is supported by the Seneca Foundation, Regional Agency for Science and Technology, Murcia, Spain (15302/PHCS/10) and Ministry of Science and Innovation, Spain (PSI11560–2009). We particularly wish to gratefully acknowledge the sample volunteers and the BioBank IBSP-CV (PT13/0010/0064) integrated in the Spanish National Biobanks Network and Valencian Biobanking Network for their collaboration.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102743.

References

[1] E.R. Gibney, C.M. Nolan, Epigenetics and gene expression, Hered. (Edinb.). Nat. Publ. Group 105 (1) (2010) 4–13.
[2] A. Riggs, V. Russo, R. Martienssen, Epigenetic Mechanisms Of Gene Regulation, Cold Spring Harbor Laboratory Press,, Plainview, N.Y, 1996.
[3] A. Saha, J. Wittmeyer, B.R. Cairns, Chromatin remodelling: the industrial revolution of DNA around histones, Nat. Rev. Mol. Cell Biol. 7 (6) (2006) 437–447.
[4] B.D. Strahl, C.D. Allis, The language of covalent histone modifications, Nature 403 (6765) (2000) 41–45.
[5] J.T. Lee, Epigenetic regulation by long noncoding RNAs, Science (80-) 338 (6113) (2012) 1435–1439.
[6] P.A. Jones, Functions of DNA methylation: islands, start sites, gene bodies and beyond, Nat. Rev. Genet. Nat. Publ. Group 13 (7) (2012) 484–492.
[7] Z.D. Smith, A. Meissner, DNA methylation: roles in mammalian development, Nat. Rev. Genet 14 (3) (2013) 204–220.
[8] D. Schübeler, Function and information content of DNA methylation, Nature 517 (7534) (2015) 321–326.
[9] V.K. Rakyan, T.A. Down, S. Maslau, T. Andrew, T.P. Yang, H. Beyan, et al., Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains, Genome Res 20 (4) (2010) 434–439.
[10] S. Bocklandt, W. Lin, M. Sehl, F. Sánchez, J. Sinsheimer, S. Horvath, et al., Epigenetic Predictor of Age, PLoS One 6 (6) (2011), e14821.
[11] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, et al., Methylation of ELOVL2 gene as a new epigenetic marker of age, Aging Cell 11 (6) (2012) 1132–1134.
[12] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, et al., Genome-wide methylation profiles reveal quantitative views of human aging rates, Mol. Cell 49 (2) (2013) 359–367.
[13] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (10) (2013) R115.
[14] Å. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, PLoS One 8 (6) (2013), e67378.
[15] I. Florath, K. Butterbach, H. Müller, M. Bewerunge-hudler, H. Brenner, 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 23 (5) (2014) 1186–1201.
[16] H. Alsaleh, P.R. Haddrill, Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip. Forensic Sci Int, 303, Elsevier Ireland Ltd,, 2019, 109944.
[17] S.K. Merid, A. Novoloaca, G.C. Sharp, L.K. Küpers, A.T. Kho, R. Roy, et al., Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age, Genome Med. Genome Med. 12 (1) (2020) 1–17.
[18] M. Jung, G.P. Pfeifer, Aging and DNA methylation, BMC Biol. 13 (1) (2015) 7.
[19] A. Freire-Aradas, C. Phillips, M.V. Lareu, Forensic individual age estimation with DNA: From initial approaches to methylation tests, Forensic Sci. Rev. 29 (2) (2017) 121–144.
[20] M.E. Levine, A.T. Lu, A. Quach, B.H. Chen, T.L. Assimes, S. Bandinelli, et al., An epigenetic biomarker of aging for lifespan and healthspan, Aging (Albany NY) 10 (4) (2018) 573–591.
[21] A.T. Lu, A. Quach, J.G. Wilson, A.P. Reiner, A. Aviv, K. Raj, et al., DNA methylation GrimAge strongly predicts lifespan and healthspan, Aging (Albany NY) 11 (2) (2019) 303–327.
[22] R. Noroozi, S. Ghafouri-Fard, A. Pisarek, J. Rudnicka, M. Spólnicka, W. Branicki, et al., DNA methylation-based age clocks: From age prediction to age reversion, Ageing Res Rev. 68 (2021), 101314.
[23] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J.M. Le, et al., High density DNA methylation array with single CpG site resolution. Genomics, 98, Elsevier Inc,, 2011, pp. 288–295.
[24] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al., Aging of blood can be tracked by DNA methylation changes at just three CpG sites, Genome Biol. 15 (2) (2014) R24.
[25] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Z. Makowska, A. Pałeczka, et al., Development of a forensically useful age prediction method based on DNA methylation analysis, Forensic Sci. Int Genet 17 (2015) 173–179.
[26] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, 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 24 (2016) 65–74.
[27] 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.
[28] 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, Elsevier,, 2019, pp. 1–8.
[29] 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.
[30] 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 (10) (2015) 922–930.
[31] H.Y. Lee, S.R. Hong, J.E. Lee, I.K. Hwang, N.Y. Kim, J.M. Lee, et al., Epigenetic age signatures in bones, Forensic Sci. Int. Genet. (2020) 46.
[32] J. Fleckhaus, P.M. Schneider, Novel multiplex strategy for DNA methylation-based age prediction from small amounts of DNA via Pyrosequencing, Forensic Sci. Int. Genet. 44 (2020), 102189.
[33] D. Zubakov, F. Liu, I. Kokmeijer, Y. Choi, J.B.J. van Meurs, W.F.J. van IJcken, et al., Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length, Forensic Sci. Int. Genet. 24 (2016) 33–43.
[34] 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, in: Forensic Sci. Int. Genet., Elsevier,, 2018, pp. 215–226.
[35] A. Heidegger, C. Xavier, H. Niederstätter, M. de la Puente, E. Pośpiech, A. Pisarek, et al., Development and optimization of the VISAGE basic prototype tool for forensic age estimation, in: Forensic Sci. Int. Genet., Elsevier,, 2020, 102322.
[36] A. Woźniak, A. Heidegger, D. Piniewska-róg, E. Pośpiech, A. Pisarek, E. Kartasińska, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, Buccal Cells Bones 13 (5) (2021).
[37] A. Aliferi, S. Sundaram, D. Ballard, A. Freire-Aradas, C. Phillips, M.V. Lareu, et al., Combining current knowledge on DNA methylation-based age estimation towards the development of a superior forensic DNA intelligence tool, Forensic Sci. Int. Genet. (2021), 102637.
[38] I. Smeers, R. Decorte, W. Van de Voorde, B. Bekaert, Evaluation of three statistical prediction models for forensic age prediction based on DNA methylation, in: Forensic Sci. Int. Genet., Elsevier, 2018, pp. 128–133.
[39] C. Xu, H. Qu, G. Wang, B. Xie, Y. Shi, Y. Yang, et al., A novel strategy for forensic age prediction by DNA methylation and support vector regression model, Sci. Rep. 5 (2015) 17788.
[40] A. Vidaki, D. Ballard, A. Aliferi, T.H. Miller, L.P. Barron, D. Syndercombe Court, DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, Forensic Sci. Int Genet 28 (2017) 225–236.
[41] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. 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.
[42] A. Freire-Aradas, C. Phillips, L. Girón-Santamaría, A. Mosquera-Miguel, A. Gómez-Tato, M.Á. Casares de Cal, et al., Tracking age-correlated DNA methylation markers in the young, in: Forensic Sci Int Genet, Elsevier, 2018, pp. 50–59.
[43] R.S. Alisch, B.G. Barwick, P. Chopra, L.K. Myrick, G.A. Satten, K.N. Conneely, et al., Age-associated DNA methylation in pediatric populations, Genome Res. 22 (4) (2012) 623–632.
[44] M.F. Fraga, E. Ballestar, M.F. Paz, S. Ropero, F. Setien, M.L. Ballestar, et al., Epigenetic differences arise during the lifetime of monozygotic twins, Proc. Natl. Acad. Sci. USA 102 (30) (2005) 10604–10609.
[45] 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.
[46] Ehrich M., Correll D., Boom D.Van Den Introduction to EpiTYPER for quantitative DNA methylation analysis using the MassARRAY ® System. Seq Appl Note [Internet]. 2006;Doc. No. 8(8876):1–8. Available from: www.sequenom.com.

[47] Koenker R., Portnoy S., Ng P.T., Zeileis A., Grosjean P., Moler C., et al. Quantile Regression, Package " quantreg." 2019.
[48] Cannon A. Package " qrnn." 2019.
[49] Steinwart I., Thomann P., Farooq M. Package " liquidSVM." 2017.
[50] Alfons A. Package "cvTools": Cross-validation tools for regression models. 2015.
[51] Wickham H., Chang W. Create Elegant Data Visualisations Using the Grammar of Graphics, Package "ggplot2." 2019.
[52] M. Eipel, F. Mayer, T. Arent, M.R.P. Ferreira, C. Birkhofer, U. Gerstenmaier, et al., Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures, Aging (Albany NY) 8 (5) (2016) 1034–1048.
[53] M. Spólnicka, E. Pośpiech, B. Pepłońska, R. Zbieć-Piekarska, Makowska, A. Pięta, et al., DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups, Int. J. Leg. Med. 132 (1) (2018) 1–11.

Artículo 2: Epigenetic age estimation in saliva and in buccal cells

A. Ambroa-Conde, L. Girón Santamaría, A. Mosquera-Miguel, C. Phillips, M. A. Casares de Cal, A. Gómez-Tato, J. Álvarez-Dios, M. de la Puente, J. Ruiz-Ramírez, M. V. Lareu, A. Freire-Aradas.

Forensic Science International: Genetics, volumen 61, página 102770, 2022, PMID: 36057238, DOI: 10.1016/j.fsigen.2022.102770

Forensic Science International: Genetics 61 (2022) 102770

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Epigenetic age estimation in saliva and in buccal cells

A. Ambroa-Conde [a], L. Girón-Santamaría [a], A. Mosquera-Miguel [a], C. Phillips [a], M.A. Casares de Cal [b], A. Gómez-Tato [b], J. Álvarez-Dios [c], M. de la Puente [a], J. Ruiz-Ramírez [a], M.V. Lareu [a], A. Freire-Aradas [a,*]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain
[c] Faculty of Mathematics, University of Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
DNA methylation
Forensic age estimation
Logistic regression
Quantile regression
SNaPshot
Saliva
Buccal swab
Buccal cells

ABSTRACT

Age estimation based on epigenetic markers is a DNA intelligence tool with the potential to provide relevant information for criminal investigations, as well as to improve the inference of age-dependent physical characteristics such as male pattern baldness or hair color. Age prediction models have been developed based on different tissues, including saliva and buccal cells, which show different methylation patterns as they are composed of different cell populations. On many occasions in a criminal investigation, the origin of a sample or the proportion of tissues is not known with certainty, for example the provenance of cigarette butts, so use of combined models can provide lower prediction errors.

In the present study, two tissue-specific and seven age-correlated CpG sites were selected from publicly available data from the Illumina HumanMethylation 450 BeadChip and bibliographic searches, to help build a tissue-dependent, and an age-prediction model, respectively. For the development of both models, a total of 184 samples (N = 91 saliva and N = 93 buccal cells) ranging from 21 to 86 years old were used. Validation of the models was performed using either k-fold cross-validation and an additional set of 184 samples (N = 93 saliva and N = 91 buccal cells, 21–86 years old).

The tissue prediction model was developed using two CpG sites (*HUNK* and *RUNX1*) based on logistic regression that produced a correct classification rate for saliva and buccal swab samples of 88.59 % for the training set, and 83.69 % for the testing set. Despite these high success rates, a combined age prediction model was developed covering both saliva and buccal cells, using seven CpG sites (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*) based on multivariate quantile regression giving a median absolute error (MAE): ± 3.54 years and a correct classification rate (%CP±PI) of 76.08 % for the training set, and an MAE of ± 3.66 years and a %CP±PI of 71.19 % for the testing set. The addition of tissue-of origin as a co-variate to the model was assessed, but no improvement was detected in age predictions. Finally, considering the limitations usually faced by forensic DNA analyses, the robustness of the model and the minimum recommended amount of input DNA for bisulfite conversion were evaluated, considering up to 10 ng of genomic DNA for reproducible results. The final multivariate quantile regression age predictor based on the models we developed has been placed in the open-access *Snipper* forensic classification website.

1. Introduction

Age estimation can provide key information in criminal, legal and anthropological investigations [1]. In cases where there are no suspects and the DNA profiles recovered from forensic biological samples do not match with any profile stored in national DNA databases, age prediction can play an important role guiding police investigations, which can reduce the number of potential suspects [2]. Age estimation may also improve the prediction of phenotypic characteristics related to aging, e.g. hair colour [3] or male pattern baldness [4]. Additionally, if the prediction models develop enough accuracy, legal disputes could potentially be supported by age estimation [5]. In all these cases, chronological age rather than biological age needs to be inferred [6].

DNA methylation has become the gold standard biomarker for

* Corresponding author.
E-mail address: ana.freire@usc.es (A. Freire-Aradas).

https://doi.org/10.1016/j.fsigen.2022.102770
Received 23 June 2022; Received in revised form 22 August 2022; Accepted 24 August 2022
Available online 27 August 2022

human age estimation. This epigenetic signature consists of the addition of a methyl group (-CH_3) to the 5′ carbon of cytosines positioned next to guanines (CpG nucleotides) [7]. Age correlation with DNA methylation has been largely confirmed by a broad range of epigenetic studies [8–16]. Based on the DNA methylation values of age correlated CpG sites, multiple forensic age prediction models have been developed to date, reviewed in [1]. Since DNA methylation is tissue-specific [17], most of these epigenetic clocks have been based on specific forensic tissues, including blood [18–21], buccal swabs [22–24], saliva [23,25] and semen [26,27]. More recently, skeletal remains, e.g., bones and teeth have been studied [28,29].

Whole blood is not uniformly composed of identical cell types, but consists of distinct cell populations in varying proportions. As methylation profiles of peripheral blood mononuclear cells and granulocytes have been identified [30], cell heterogeneity could act as a confounder. However, studies have observed that DNA methylation for age correlated CpG sites does not vary significantly across sorted blood cells from healthy subjects [13], and subsequently, most forensic age prediction models were based on whole blood treated as a homogeneous tissue.

Another tissue source that lacks cellular homogeneity is the oral cavity, where saliva and buccal swabs have different varied proportions of leucocytes and epithelial cells [31]. This difference in cell content could potentially create differences in DNA methylation for specific CpG sites, and this phenomenon was previously observed for *ELOVL2* and *FHL2* [23], indicating that both sample types cannot be considered a single biological source beforehand.

Nevertheless, considering that deconvolution to assign the specific biological source - saliva or buccal swabs - to forensic oral cavity specimens is difficult to achieve, e.g., cigarette butts, the development of a single age prediction model covering both tissues represents a practical approach.

A similar approach has already been proposed by Horvath et al. [32], developing the “skin & blood clock”, an epigenetic clock based on 391 CpGs that covers samples originating from blood, skin, saliva, buccal cells, as well as from four additional somatic tissues. The age prediction model reported by Jung et al., is more focused on forensic specimens, and is based on 5 CpG sites applicable to either blood, saliva or buccal cells [23].

In the present study, we focused on specimens from the oral cavity aiming to develop a tissue prediction model that can differentiate saliva from buccal cells, as well as an age prediction model covering both tissues, since most forensic samples related to the oral cavity will comprise a mixture of saliva and buccal cells. Additionally, to include the tissue-of-origin as a co-variable do not improve age predictions. Selection of candidate tissue-specific and age correlated CpG sites was based on the assessment of public data from Illumina HumanMethylation 450 K. Then, 184 volunteers (21–86 years old) were analyzed using SNaPshot™, after collection of either saliva and buccal swabs from the same individual (N = 368). A proportion of the analyzed samples were used to develop the training set (N = 184), while an additional part was used as a testing set for model validation purposes (N = 184). As a result, a tissue prediction model (saliva vs buccal cells) using logistic regression and based on 2 CpG sites was developed. In parallel, an age prediction model covering these tissues together and based on multivariate quantile regression analysis was developed for 7 CpG sites showing the highest correlation with age. Since SNaPshot™ needs a preliminary step of bisulfite conversion that degrades the DNA, requiring high levels of input DNA, we made an evaluation of serial dilutions with this detection system to determine the limits of the assay.

2. Material and methods

2.1. Samples, DNA extraction and quantification

A total of 368 samples, 184 total saliva and 184 buccal cells, were collected from 184 healthy Spanish volunteers from 21 to 86 years old. Based on this set of samples, for the saliva-specific and buccal swab-specific age prediction models, the whole set of 184 saliva and 184 buccal swabs, respectively, were directly used as training sets. For the tissue-combined age prediction model, a random selection was made to generate training and test sets balanced in terms of sample size, distribution of ages and represented tissues. Each group had 184 individuals with the full age range 21–86 years. The training set consisted of 91 saliva and 93 buccal cell samples, while the testing group had 93 saliva and 91 buccal cell samples.

All samples were taken with written informed consent obtained from the donors. Ethical approval was obtained from the ethics committee of investigation in Galicia, Spain (CAEI: 2013/543). Buccal swabs were air-dried and stored at room temperature and total saliva was collected with 15 mL falcon tubes and frozen at − 20 °C until DNA extraction. Genomic DNA was extracted from the whole swab and from 500 μL of total saliva with phenol/chloroform extraction [33]. All DNA samples were quantified by Qubit® dsDNA High Sensitivity (HS) or dsDNA Broad Range (BR) Assay kits (Thermo Fisher) following manufacturer’s guidelines.

2.2. CpG site selection

Selection of candidate CpG sites was based on both bibliographic searches as well as statistical assessment of NCBI GEO methylation studies using public data from the Illumina HumanMethylation450KBeadChip. Tissue-specific CpG site selection was based on the statistical assessment of the methylation β-values from GSE48472 [34] (blood, saliva and buccal cells). To check for absence of correlation with age for the selected tissue-specific markers; GSE87571 [14] GSE92767 [25] and GSE50586 [35] were used. Furthermore, the bibliographic review was focused on publications from 2011 to 2019, and searched for markers presenting a high correlation with age in different tissues: blood [18,20,28,36], saliva [9,37], and buccal cells [22,38]. Additionally, methylation β-values from GSE92767 [25] were statistically assessed to seek to identify additional age-correlated CpG sites.

2.3. Primer design

The flanking regions of the selected CpGs were screened using the UCSC genome browser (https://genome.ucsc.edu/) for the current human genome assembly (GRCh38/hg38), covering 150 bp upstream and downstream of the target CpG. The PCR primer and Single Base Extension (SBE) primer designs were made using BatchPrimer 3 v1.0 [39] applying the following parameters for PCR primers: optimal melting temperature 58 °C, optimal primer length 20 bp and optimal amplicon length 90 bp; and for the SBE primer design: optimal melting temperature 50 °C and optimal probe length 20 bp. Poly-CT tails were added to the SBE primers for size separation.

2.4. Bisulfite conversion, PCR conditions and purification of PCR products

Bisulfite conversion of 100 ng of extracted genomic DNA was carried out with the MethylEdge™ Bisulfite Conversion System (Promega) following manufacturer’s guidelines, obtaining an elution volume of 20 μL. A PCR multiplex amplification in 10.7 μL reaction volume adding 1.5 μL of converted DNA was carried out using 0.3 μL of 250 U AmpliTaq Gold™ DNA Polymerase, 1.5 μL of 10X Buffer II, 3.9 μL of 25 mM $MgCl_2$ (all from Applied Biosystems, AB), 1.5 μL of 32 ng/μL bovine serum albumin, 1 μL of 10 mM GeneAmp® dNTP Mix with dTTP (AB) and 1 μL of primer mix (0.083–5 μM of each primer, Metabion International). PCR cycling used a GeneAmp® PCR system 2720 (AB) with cycling conditions: 95ºC for 11 min; 34 cycles of 94ºC for 20 s, 56ºC for 60 s and 72ºC for 30 s, and a final extension of 72ºC for 7 min

After checking amplification yields in 1 % agarose gels, a purification of 2.5 μL of PCR product was performed adding 1 μL of ExoSAP-IT™ PCR

Product Cleanup Reagent (AB) at 37 °C for 45 min and 80 °C for 15 min

2.5. Single base extension and capillary electrophoresis

Multiplex SBE reactions were performed in a total volume of 6 μL using 2 μL of purified PCR product, 2.5 μL of SNaPshot™ kit (AB) and 1.5 μL of SBE primers (0.51–6 μM of each primer, Metabion International) with cycling conditions: 30 cycles of 96 °C for 10 s, 55 ° for 5 s, and 60 °C for 30 s

After the SNaPshot reaction, extension products were purified by adding 1 μL of Shrimp Alkaline Phosphatase Recombinant (AB) to the total SNaPshot reaction and incubating at 37 °C for 80 min with inactivation at 85 °C for 15 min

Capillary electrophoresis was performed with an ABI3130xl Genetic Analyzer (AB) using 0.1 μL of GeneScan™ 120 LIZ™ dye Size Standard (Thermo Fisher) and 10 μL of HiDi™ Formamide (AB) per sample, adding 9.5 μL of load mix and 1.5 μL of purified SNaPshot product. Results were analyzed with GeneMapperID v3.2 (AB) and the DNA methylation level at each CpG was calculated by dividing the height of the methylated peak by the sum of the heights of the methylated and unmethylated peaks. The latter values were multiplied by a correction factor of 2, when working with reverse primers and 1.6 for forward primers, to overcome differences at fluorochrome signal intensities.

2.6. Statistical analyses

All samples were run in duplicate. The average of the DNA methylation levels in both replicates was used for the statistical analyses. Correlations between age and DNA methylation levels were evaluated using the Spearman Correlation test (r_s). To analyze the reproducibility of the dilutions and the inter-individual variability, the standard deviation (SD) was used (threshold SD > 0.1). Normality was assessed using the Shapiro-Wilk test applied to the residuals of the independent linear regression models tested for each CpG (p-value < 0.05). Logistic regression was used to develop the tissue prediction model using the *pROC* R package [40]. A multivariate quantile regression model was used to build the age prediction model using the *quantreg* R package [41]. Cross-validation of the prediction models was performed with a k-fold cross-validation (k = 10) using the *cvTools* R package [42]. The corresponding predictive accuracy was measured with the following performance metrics: sensitivity, specificity, area under the curve (AUC) and percentage of correct classifications for tissue prediction; and the median absolute error (MAE), the mean absolute error (MAE_{mean}), the root-mean-square error (RMSE) and percentage of correct classifications within the prediction intervals (%CP±PI) for age prediction. The representation of predicted versus chronological age was made using the *ggplot2* R package [43]. All statistical analyses were carried out using R software v.4.0.3 [44] with scripts developed in-house. The sensitivity analysis was carried out using input DNA quantities for bisulfite conversion of 100 ng, 75 ng, 50 ng, 25 ng, 10 ng and 1 ng.

3. Results

3.1. Selection of candidate CpGs

The selection of candidate CpGs was divided into tissue-specific CpGs and age-correlated CpGs.

For selection of tissue-specific CpGs, the GSE48472 dataset was assessed [34]. From this dataset, samples from saliva (N = 5), buccal cells (N = 5) and blood (N = 5) were selected and differences in the corresponding DNA methylation values calculated. A total of 17 CpG sites with the highest differences in DNA methylations levels were found (Table 1): 5 CpGs presenting the highest differences between blood and buccal cells (>|0.72|); 6 CpGs between blood and saliva (>|0.45|) and 6 CpGs between saliva and buccal cells (≥|0.5|).

Once the markers had been selected, absence of correlation with age was evaluated using the following datasets: GSE92767 (saliva) [25], GSE50586 (buccal cells) [35] and GSE87571 (blood) [14]. From the 17 selected tissue-specific CpGs, three displayed correlations with age (r_s >|0.5|): cg01680010 (r_s =−0.607) in buccal cells and cg13408086 (r_s =0.609) and cg08466792 (r_s =0.575) in blood, so were discarded. Based on these results, one CpG site per tissue combination was selected. This selection was initially based on the highest difference displayed by the DNA methylation values observed in pairs of tissues. However, several failures in PCR primer design led to a final selection of cg04915566 (*RUNX1*) for blood-buccal cells, cg16606773 (*RIN2*) for blood-saliva and cg03044684 (*HUNK*) for saliva-buccal cells.

Selection of age-correlated CpGs was based on the assessment of DNA methylation values from GSE92767 (saliva samples, N = 54, 18–73 years old) [25]. In order to select the method to be used for marker selection, normality was evaluated for GSE92767 data, obtaining that 15 % of the residuals of the models (independent linear regression models for each CpG) presented a lack of normality (p-value < 0.05), therefore, the Spearman test was used. For this analysis, CpG sites presenting a Spearman correlation coefficient equal to or greater than |0.8| were selected, providing 49 CpG sites correlated with age (Supplementary Table S1). From this preliminary set of sites, those CpGs with a minimum difference of 0.3 between the highest and lowest methylation

Table 1
Summary of the 17 selected tissue-specific CpG sites based on the statistical assessment of GSE48472. CpG sites correlated with age (r_s >|0.5|) are marked in bold.

Tissue's comparison	Gene	CpG_ID	GRCh38 chromosome position	Differences between pairs of tissues	Correlation with age (r_s blood)	Correlation with age (r_s buccal cells)	Correlation with age (r_s saliva)
Blood-Buccal cells	*RUNX1*	cg04915566	chr21:35049175	0.723	0.023	0.006	-0.309
	MAML2	cg08141395	chr11:96254218	0.748	0.006	0.043	-0.306
	RGS1	cg10861751	chr1:192575586	0.733	-0.024	0.055	-0.266
	EXD3	cg13408086	chr9:137326945	0.724	**0.609**	-0.337	-0.481
	NCKAP1L	cg16509569	chr12:54497850	0.721	-0.019	-0.190	-0.318
Blood-Saliva	*CDC25B*	cg02737268	chr20:3799535	0.483	0.287	-0.079	0.439
	DOT1L	cg04173586	chr19:2167497	0.459	0.001	0.129	0.408
	RIN3	cg15443535	chr14:92687972	0.476	-0.152	0.411	0.399
	none	cg16149628	chr11:1771344	0.471	0.023	0.166	0.270
	RIN2	cg16606773	chr20:19975162	0.459	-0.179	-0.153	-0.049
	WDFY1	cg23363263	chr2:223887272	0.452	-0.102	-0.043	0.218
Saliva-Buccal cells	none	cg01680010	chr7:97017805	0.500	0.148	**-0.607**	-0.079
	none	cg02939659	chr14:101587733	0.500	-0.048	0.472	0.389
	HUNK	cg03044684	chr21:31875719	0.503	-0.065	0.055	0.247
	PAX9	cg07459252	chr14:36661007	0.502	0.334	-0.104	-0.106
	none	cg08466792	chr5:3603113	0.512	**0.575**	-0.362	-0.378
	SIM2	cg25446076	chr21:36710849	0.523	0.379	-0.349	-0.332

values were selected, to give ten candidate CpGs (Table 2) for age prediction in saliva.

As the statistical analysis for selection of age correlated CpG sites was based on saliva samples, but the study also covered buccal cells; a bibliographic search to find genes that show correlation with age in additional somatic tissues was carried out. In the reviewed publications, certain markers were repeatedly found to correlate with age in the tissues of interest (saliva, buccal cells and blood): *PDE4C* [18,20,22,28,37], *EDARADD* [9,28,38] and *ASPA* [18,20,22,28,36,37], and therefore these genes were the focus of further evaluation in our study (included in Table 2).

3.2. Development of an optimized multiplex

From the above analyses, 16 markers were selected: 3 tissue-specific CpG sites (*RUNX1*, *RIN2* and *HUNK*) and 13 age-correlated CpG sites (*OTUD7A*, *FHL2*, *TRIM59*, *RHBDL2*, cg10501210, cg10804656, *LHFPL4*, cg13327545, *ELOVL2*, *HOXC4*, *PDE4C*, *EDARADD* and *ASPA*). PCR and SBE primers were successfully designed for the selected tissue-specific markers, and 11 age-correlated CpGs, with *RHBDL2* and cg10804656 discarded from subsequent analyses. A summary of PCR and SBE primer information is outlined in Supplementary Table S2.

First, each marker was analyzed in singleplex to check for individual amplification performance. Once this initial step was accomplished, a multiplex covering all 14 CpGs was optimized. Markers *TRIM59* and cg13327545 were not amplified in multiplex due to non-specific hybridizations leading to the final optimized multiplex of *RUNX1*, *RIN2*, *HUNK* tissue-specific markers plus *OTUD7A*, *FHL2*, cg10501210, *LHFPL4*, *ELOVL2*, *HOXC4*, *PDE4C*, *EDARADD*, *ASPA* age-correlated CpG sites. An example SNaPshot electropherogram of the optimized multiplex is shown in Fig. 1.

To check tissue-specificity and age correlation of the optimized marker set, a preliminary analysis using both saliva and buccal cells from two individuals of age extremes (23 and 86 years old) was completed (Supplementary Table S3). Tissue specificity was not detected for *RIN2*, since the same absence of methylation pattern was observed for both saliva and buccal cells. To check that the absence of methylation was not a technical problem, and since *RIN2* was selected for detecting differences between blood and saliva, blood samples from the same individuals were tested and detected DNA methylation levels of 0.39 and 0.4, respectively.

In the case of *RUNX1*, some dispersion was detected in the patterns displayed by age or tissue, preventing objective interpretation with this marker. However, *HUNK* had differences in average DNA methylation levels between both tissues (0.31 and 0.17 for saliva and buccal cells, respectively).

For age-correlation, six of the nine CpG sites (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *ASPA* and *EDARADD*) gave average DNA methylation difference between extreme ages equal to or higher than 0.19. Differences were displayed by *HOXC4*, *OTUD7A* and *FHL2* at lower levels (0.1, 0.05 and 0.12, respectively).

Table 2
Summary of the ten selected CpG sites correlated with age in saliva, based on the statistical assessment of GSE92767, as well as the 3 selected age correlated CpG sites in somatic tissues based on bibliographic review.

Gene	CpG_ID	GRCh38 chromosome position	Correlation with age (r_s)	Methylation differences at extreme ages
GSE92767 assessment				
OTUD7A	cg04875128	chr15:31483692	0.860	0.312
FHL2	cg06639320	chr2:105399282	0.824	0.322
TRIM59	cg07553761	chr3:160450189	0.803	0.305
RHBDL2	cg10500653	chr1:38941979	0.814	0.334
none	cg10501210	chr1:207823675	-0.864	0.674
none	cg10804656	chr10:22334531	0.828	0.317
LHFPL4	cg11084334	chr3:9552580	0.846	0.345
none	cg13327545	chr10:22334619	0.818	0.302
ELOVL2	cg16867657	chr6:11044644	0.898	0.385
HOXC4	cg18473521	chr12:54054481	0.824	0.389
Bibliographic review				
PDE4C	none	chr19:18233131	na	na
EDARADD	cg09809672	chr1:236394382	na	na
ASPA	cg02228185	chr17:3476273	na	na

3.3. A statistical tissue prediction model

The training set comprising 91 saliva samples and 93 buccal swabs was analyzed with the optimized multiplex to develop a tissue prediction model for saliva and buccal cells. The corresponding dispersion diagrams for *HUNK* and *RUNX1* markers is shown in Fig. 2. Dispersion correlated with the tissue-of-origin is observed, with higher methylation levels for *HUNK* in saliva samples, and for *RUNX1* in buccal cells.

In order to predict tissue of origin, logistic regression was applied exploring three different models: model 1 (*HUNK* plus *RUNX1*), model 2 (*HUNK*) and model 3 (*RUNX1*). The corresponding performance metrics are described in Table 3. Comparable AUC values of 0.95, 0.95 and 0.92 for model 1, 2 and 3, respectively were obtained. Similar percentage of correct classifications was also recorded, with model 1 having the highest value at 88.6 %. However, some differences were found with sensitivity and specificity values, considering buccal cell samples as the control (i.e., a high specificity indicates good classification of buccal cell samples and a high sensitivity good classification of saliva samples). Model 1 gave a higher sensitivity (0.96) compared to specificity (0.82). Therefore, model 1 results show that saliva samples classify better than swab samples. In contrast, model 2 gave a sensitivity of 0.78 and a specificity of 0.96; model 3 gave a sensitivity of 0.81 and a specificity of 0.9. Therefore, single marker models classify saliva samples less efficiently than buccal swab samples.

Considering the highest rate of correct classifications obtained (88.59 %), model 1 was selected for validation with a testing set of 184 samples (N = 93 saliva and N = 91 buccal cells. A correct tissue-of-origin prediction rate of 83.7 % for test set samples was obtained.

3.4. A statistical age prediction model for saliva and buccal swab samples

Tissue-independent as well as tissue-combined models were explored for age prediction. For the saliva-specific and buccal swab-specific age prediction models, 184 saliva and 184 buccal swab samples were used as training sets, respectively. The training set of 184 volunteers (N = 91 saliva and N = 93 buccal swabs) was used to develop the combined age prediction model for saliva and buccal cell samples. Dispersion plots in Fig. 3 indicate the patterns obtained for the cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD* markers adopted. Six markers showed hypermethylation with increased age (*LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *FHL2*); while cg10501210, *ASPA* and *EDARADD* had decreasing methylation levels with increasing age. If considering both tissues combined (saliva and buccal swabs), the highest correlation with age was found in *PDE4C* (r_s =0.806) and *LHFPL4* (r_s =0.805), followed by *ELOVL2* (r_s =0.659), *OTUD7A* (r_s =0.642), *EDARADD* (r_s =−0.572) and *HOXC4* (r_s =0.569). However, low levels of correlation were detected in cg10501210, *FHL2* and *ASPA* (r_s =−0.313, 0.198 and −0.332, respectively). At the same time, these three markers showed the highest levels of dispersion between saliva and buccal cells, (SD> 0.1). If taking into account both tissues independently, correlations followed a similar trend (Supplementary Fig. S1-S2). Whereas the highest age correlation was displayed by *LHFPL4* and *PDE4C* (r_s =0.815 and 0.832 in saliva and buccal swabs, respectively), the lowest levels of correlation were observed in cg10501210 (r_s =−0.429, −0.422), *FHL2* (r_s =0.392, 0.231) and *ASPA* (r_s =−0.521, −0.44).

Taking into account these observations, multivariate quantile regression was tested on several age prediction models consisted of different combinations of CpG sites: model 1 (9 CpGs: cg10501210,

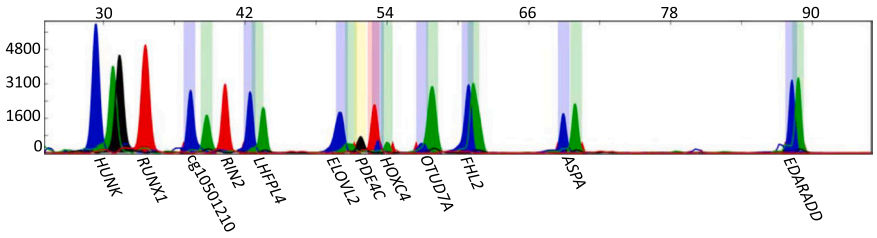


Fig. 1. Example electropherogram of the optimized SNaPshot™ multiplex assay containing 3 tissue-specific and nine age correlated CpG sites using 100 ng of genomic DNA.

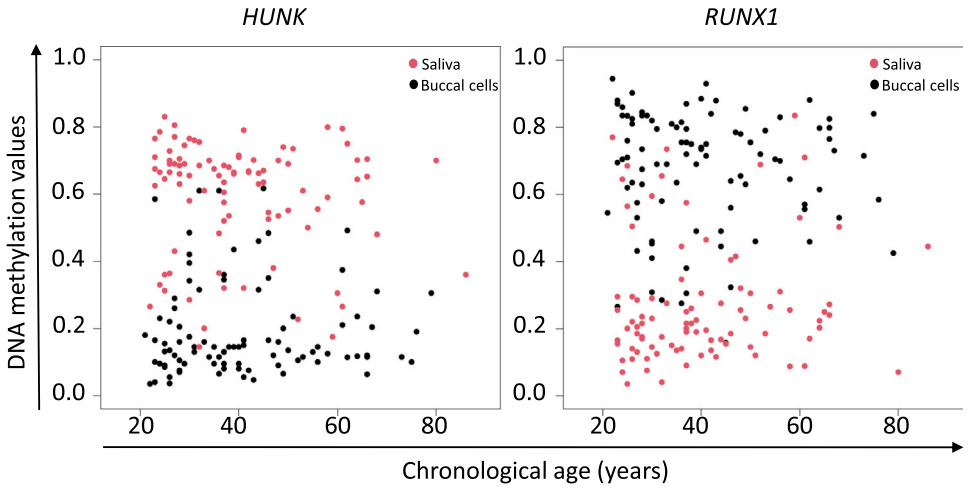


Fig. 2. Dispersion diagrams (DNA methylation values in front of chronological age) for *HUNK* and *RUNX1* (tissue-specific CpG sites) for 184 individuals from 21 to 86 years old (N = 91 saliva and N = 93 buccal swabs).

Table 3
Summary of the predictive performance metrics for the three logistic models tested on the training set (N = 91 saliva and N = 93 buccal swabs, 21–86 years old). AUC: Area under the curve.

Model	CpG_ID	Gen	AUC	Sensitivity	Specificity	Correct classifications
Model 1	cg03044684 & cg04915566	*HUNK & RUNX1*	0.95	0.96	0.82	88.59 %
Model 2	cg03044684	*HUNK*	0.95	0.78	0.96	86.87 %
Model 3	cg04915566	*RUNX1*	0.92	0.81	0.90	85.87 %

LHFPL4, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD*), model 2 (8 CpGs with *ASPA* excluded), model 3 (8 CpGs with *FHL2* excluded), model 4 (8 CpGs with cg10501210 excluded), model 5 (7 CpGs with cg10501210 and *FHL2* excluded), model 6 (7 CpGs with cg10501210 and *ASPA* excluded), model 7 (7 CpGs with *FHL2* and *ASPA* excluded) and model 8 (6 CpGs with cg10501210, *FHL2* and *ASPA* excluded). To evaluate the accuracy of the models, a k-fold cross-validation was carried out. The "k-fold" divides the total number of individuals into groups of similar sizes, in this case, 10 groups were created, each containing 10 % of the subjects. Each model was tested for each of the clusters, therefore, each time one of the clusters was selected as a test set, it faced the remaining nine that make up the training set. The corresponding performance metrics for the training sets are described in Table 4.

Inter-training set comparisons show that the correct classification rates are similar among them (%CP±PI: 76.66 %, 75.37 % and 76.23 %, for saliva, buccal swab and the combined model, respectively). However, more remarkable differences were found when comparing prediction errors, especially between the buccal swab-specific and the combined model (average MAE: ± 3.89 and ± 4.35, respectively). Nevertheless, the saliva-specific model showed a better prediction error than the combined model (average MAE: ± 3.55). Based on these results, and due to the fact that many forensic specimens will comprise a mixture of saliva and buccal cells with different cell proportions, e.g., cigarette butts, the corresponding age prediction model to be developed was selected to cover both tissues simultaneously (combined model).

Intra-training set comparisons of the combined model showed that the highest error and lowest correct classification rate were obtained with model 8 (MAE: ± 5.23, RMSE: 7.54 and %CP±PI: 74.06 %), which lacks the 3 CpG sites with the lowest levels of correlation with age and highest dispersion between saliva and buccal cells (cg10501210, *FHL2* and *ASPA*). When including these CpG sites (model 1), error decreased (MAE: ± 3.31) but the correct classification rate is only marginally improved (74.38 %). Among all models tested, the best balance between error and correct classification was obtained with model 7, which excludes *FHL2* and *ASPA* (MAE: ± 3.54, RMSE: 6.23 and %CP±PI: 76.08 %). Subsequently, we selected the age prediction model for saliva and buccal cells based on CpGs cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*. Predicted versus chronological age is plotted for the final 7-CpG age prediction model in Fig. 4. The quantiles 0.5, 0.1 and 0.9 are represented by a black line and dashed dark red lines, respectively and the gray line represents perfect correlation. The Fig. 4 plot shows that the 0.5 quantile line is more separated in older ages, possibly due to the low number of samples available for this age range. The non-parallel prediction intervals also show the reduced precision in the highest age ranges.

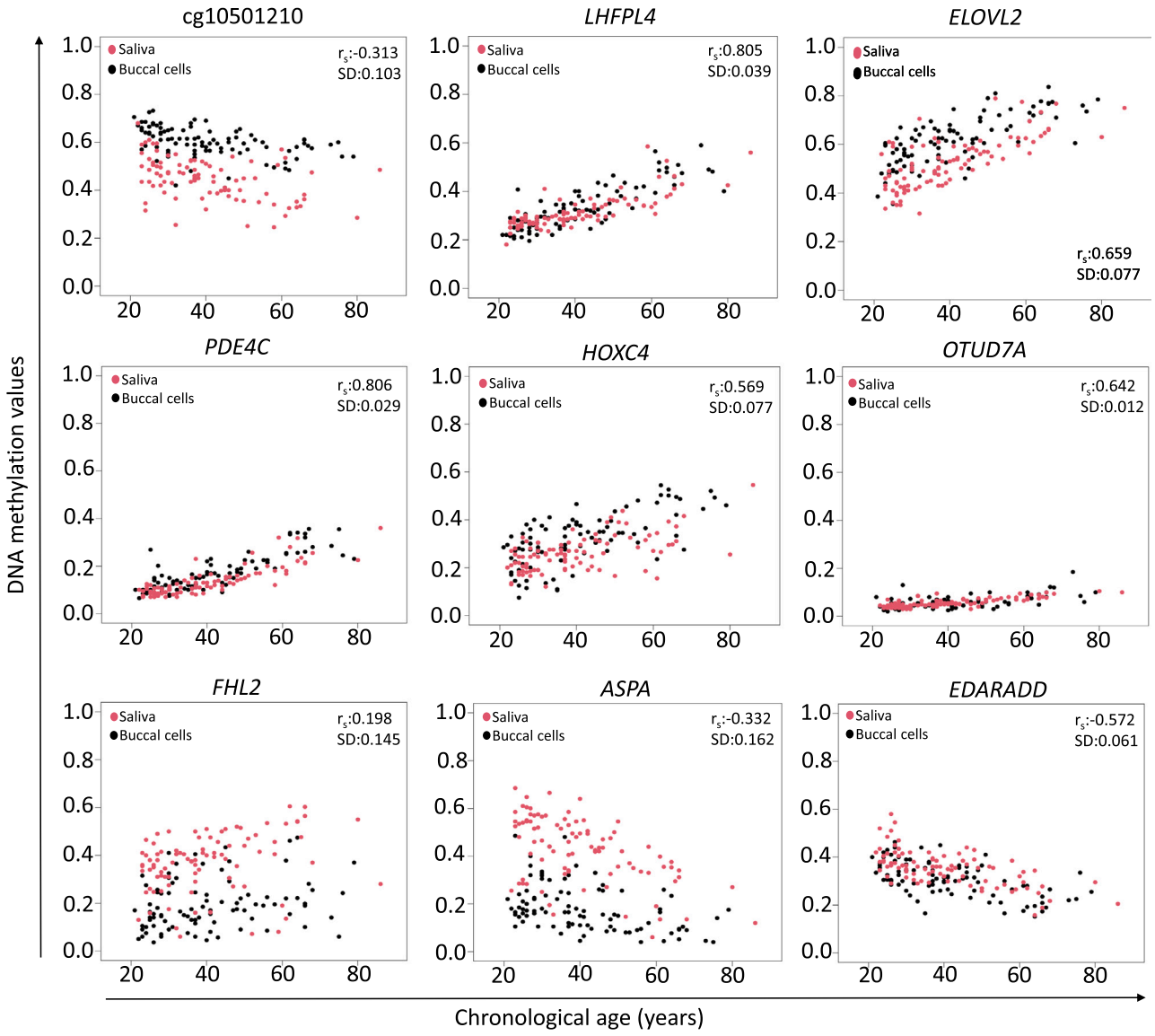


Fig. 3. Dispersion diagrams (DNA methylation values in front of chronological age) for cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD* (age correlated CpG sites) for 184 individuals from 21–86 years old (N = 91 saliva and N = 93 buccal swabs).

As well as cross-validation, an additional validation step consisted of a testing set of 184 samples (N = 93 saliva and N = 91 buccal swabs) ranging from 21–86 years old that were analyzed using the final age prediction model, providing an MAE of ± 3.66 % and 71.2 % correct classifications. The final online age prediction model developed in our study has now been placed in the open-access *Snipper* forensic classification website and is freely available at: http://mathgene.usc.es/cgi-bin/snps/age_tools/processmethylation-saliva-buccalswab.cgi. The underlying model equations for predicted age and prediction intervals are the following:

Predicted age in years = 29·33 - (50·52 x cg10501210) + (9·23 x *LHFPL4*) + (36·46 x *ELOVL2*) + (74·32 x *PDE4C*) + (11·23 x *HOXC4*) + (84·74 x *OTUD7A*) - (15·03 x *EDARADD*)·

Minimum Prediction (MinPred – q10) = 29·36 - (42·87 x cg10501210) + (15·41 x *LHFPL4*) + (11·09 x *ELOVL2*) + (74·17 x *PDE4C*) + (32·51 x *HOXC4*) + (29·13 x *OTUD7A*) - (20·54 x *EDARADD*)·

Maximum Prediction (MaxPred – q90) = 11·3 - (43·57 x cg10501210) + (20·74 x *LHFPL4*) + (54·72 x *ELOVL2*) + (78·25 x *PDE4C*) - (7·06 x *HOXC4*) + (179·95 x *OTUD7A*) + (4·16 x *EDARADD*)·

Once the age prediction model was generated, the possibility that the tissue could be considered as an additional variable was evaluated. To assess this, the prediction model was generated again by adding the tissue-of-origin of each of the samples in the training set as a co-variable. For this extended model, an MAE of ± 3.84 years, RMSE of 6.31 and % CP±PI of 78.22 % was obtained after cross-validation. Next, in order to evaluate the test set, the 2-CpG prediction model was used to predict the tissue-of-origin of the test samples. Adding the inferred tissue to the test set, produced an MAE of ± 3.78 years, RMSE of 6.6 and %CP±PI of 70.11 %. Comparing these results with those obtained when using the model without tissue source prediction, indicates the tissue as a co-variable does not improve the model.

3.5. Forensic validation of the age prediction model

To evaluate the predictive tests developed for the analysis of typical forensic samples with degradation and/or low-level DNA, the robustness and sensitivity of the final model were assessed.

A chain of models was generated by deleting one of the CpGs included in the final model, simulating random loss of one of the markers. For each of the six CpGs models generated, the training set was

Table 4
Summary of predictive performance metrics for the eight multivariate quantile regression models tested, based on three training sets: the saliva training set (N = 184 saliva, 21–86 years old), the buccal swab training set (N = 184 buccal swabs, 21–86 years old) and the combined training set (N = 91 saliva and N = 93 buccal swabs, 21–86 years old). All data represent the k-fold cross-validation. The selected model, based on the best balance between error and correct classification, is marked in bold. MAE: median absolute error, MAE_{mean}: mean absolute error, RMSE: root-mean-square error and %CP±PI: percentage of correct classifications within the prediction intervals.

Tissue	Model	CpG number	MAE	MAE_{mean}	RMSE	%CP±PI
Saliva	Model 1	9 CpGs	±3.17	±4.79	6.46	76.55 %
	Model 2	8 CpGs with *ASPA* excluded	±2.98	±4.66	6.4	77.11 %
	Model 3	8 CpGs with *FHL2* excluded	±3.29	±4.76	6.47	75.49 %
	Model 4	8 CpGs with cg10501210 excluded	±3.79	±5.04	6.76	75.59 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±3.85	±5.17	6.93	76.61 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±3.96	±4.97	6.69	74.45 %
	Model 7	7 CpGs with *FHL2* and *ASPA* excluded	±3.31	±4.69	6.37	78.74 %
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±4.02	±5.10	6.91	78.74 %
Buccal swab	Model 1	9 CpGs	±3.85	±5.01	6.35	75.47 %
	Model 2	8 CpGs with *ASPA* excluded	±4.41	±5.09	6.45	74.91 %
	Model 3	8 CpGs with *FHL2* excluded	±4.13	±4.90	6.24	75.53 %
	Model 4	8 CpGs with cg10501210 excluded	±4.45	±5.27	6.66	76.64 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±4.89	±5.52	6.94	74.42 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±4.22	±5.15	6.63	73.86 %
	Model 7	7 CpGs with *FHL2* and *ASPA* excluded	±4.16	±4.99	6.36	75.47 %
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±4.72	±5.43	6.86	76.64 %
Combined (saliva and buccal swabs)	Model 1	9 CpGs	±3.31	±4.57	6.06	74.38 %
	Model 2	8 CpGs with *ASPA* excluded	±3.66	±4.75	6.20	74.99 %
	Model 3	8 CpGs with *FHL2* excluded	±3.67	±4.78	6.32	74.36 %
	Model 4	8 CpGs with cg10501210 excluded	±3.78	±5.05	6.52	77.72 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±4.18	±5.43	6.96	77.28 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±3.77	±4.93	6.39	80.96 %
	Model 7	**7 CpGs with *FHL2* and *ASPA* excluded**	**±3.54**	**±4.79**	**6.23**	**76.08 %**
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±5.23	±5.93	7.54	74.06 %

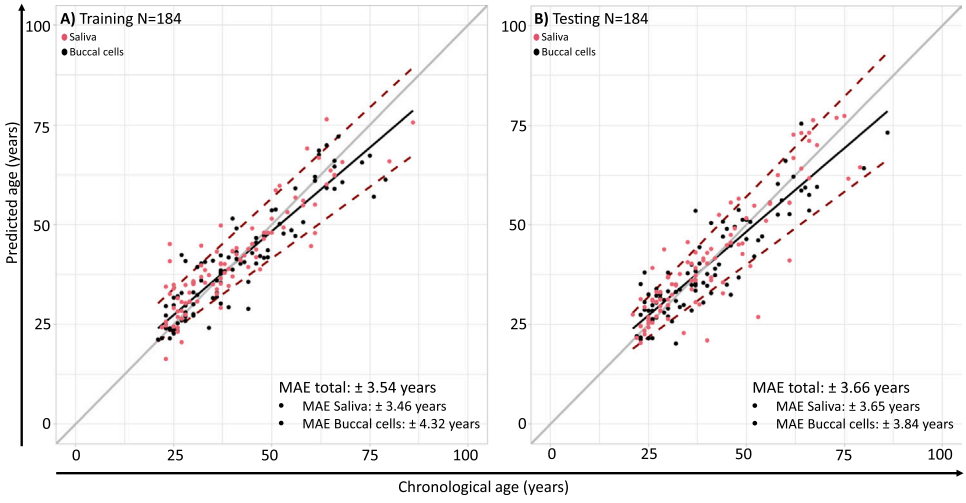


Fig. 4. Predicted versus chronological age for the final age prediction model for saliva and buccal cells for A) the training set composed of 184 individuals from 21–86 years old (N = 91 saliva and N = 93 buccal swabs) and for B) the testing set composed of 184 samples from 21–86 years old (N = 93 saliva and N = 91 buccal swabs). Predictions were performed under multivariate quantile regression using seven markers: cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*. The black diagonal line represents the 0.5 quantile and the discontinuous dark red lines the corresponding 0.1 and 0.9 quantiles. The gray line represents perfect correlation. The data represent the k-fold cross-validation.

evaluated by cross-validation and the test set. Results are outlined in Supplementary Table S4. This analysis identified those markers with the strongest contribution to the final age prediction model. The exclusion of cg10501210 increased the MAE to ± 5.23 years in the cross-validation of the training set, and exclusion of *PDE4C* increased the classification error to ± 4.03 years in the testing set. Excluding the four other markers did not greatly affect the errors obtained compared to the full 7-CpG model, so the impact of their loss is minimal.

Lastly, bisulfite conversion was performed using 100 ng of genomic DNA. To evaluate if lower quantities of input DNA could produce results of comparable quality, serial dilutions were tested on two individuals (23 and 79 years old) for both saliva and buccal cells, using input DNA quantities for bisulfite conversion of 100 ng, 75 ng, 50 ng, 25 ng, 10 ng and 1 ng. The corresponding DNA methylation values and predicted ages are listed in Supplementary Table S5. To evaluate the differences detected in DNA methylation values between input DNAs, the standard deviation (SD) was used for comparisons (Supplementary Table S6). No standard deviations higher than 0.1 were observed in any of the markers up to 10 ng. For 1 ng only 4 markers presented a higher deviation than 0.1: *ELOVL2* (SD=0.19 and SD=0.20) in two of the four samples analyzed, *RUNX1* (SD=0.20) in one sample, cg10501210 (SD=0.13) and *HOXC4* also in single samples (SD=0.16).

4. Discussion

Individual age estimation has been a topic of great interest in forensic genetics for the last years. DNA methylation has become the biomarker of choice for inferring this characteristic [45], with prediction models published using several techniques [19,20,24,38,46,47] and different tissues [18–29], although most of them have focused on blood samples. Other tissues of relevance for forensic DNA analysis, and for which age prediction models are beginning to be developed are saliva

and buccal cells [22–25]. Cellular composition of saliva and buccal swab samples has been shown to be different, with saliva composed of a majority of leukocytes and buccal swab samples of epithelial cells [31]. However, it has also been observed in previous studies that cellular proportions can vary greatly between individuals, with the saliva samples containing a variable quantity of leukocytes in the range 16–95 %, and buccal swab samples between 5 %–65 % [22,48]. Taking this into consideration, an initial step of the present study was to develop a prediction model in order to infer the tissue of origin.

The selected tissue prediction markers have not been previously reported. Each marker was selected considering differences between pairs of tissues: saliva versus buccal cells (*HUNK*), blood versus buccal cells (*RUNX1*) and blood versus saliva (*RIN2*). From these three candidate markers, *RIN2* showed no variation in the DNA methylation patterns for the tissues of interest (saliva and buccal swabs) and therefore, was discarded from subsequent analyses. In contrast, differences were distinct for *RUNX1* and particularly *HUNK* (Fig. 2), with this pair showing opposite trends in DNA methylation levels (average DNA methylation: 0.576 for saliva versus 0.199 for buccal cells in *HUNK*, and 0.297 for saliva versus 0.670 for buccal cells in *RUNX1*). Likely due to the possible variations in the composition of the tissues collected (higher percentages of leukocytes or epithelial cells) [22], in some samples, differences were also observed within the same tissue, for example *RUNX1* gave differences up to 0.34 between some saliva samples. Additional cell-specific markers such as *CD6, SERPINB5* [22] and *PTPN7* [25] have been reported in other studies. The selection of these different markers could be due to screens made of alternative datasets. For the selection of *CD6* and *SERPINB5,* Eipel et. al used datasets GSE50586 [35] and GSE39981 [49], the former with data from buccal swab samples and the latter from blood samples. Using these data in combination, they selected CpGs that showed differences according to the tissue of origin. It should be noted that in our case we only used GSE50586 to evaluate whether tissue-specific markers related to buccal cells were correlated with age. On the other hand, the selection of *PTPN7* came from the Hong et al. study evaluating DNA methylation differences between blood and buccal cells. In our case, we selected a dataset containing samples of different tissues for each individual [34], trying to limit the possible differences between individuals related to the varied cellular proportions in saliva and buccal swab samples.

HUNK (hormonally up-regulated Neu-associated kinase) is a gene predicted to be involved in intracellular signal transduction and protein phosphorylation, while the protein encoded by *RUNX1* (*RUNX* family transcription factor 1) is involved in the development of normal hematopoiesis. Once both genes were selected as candidate markers for the inference of the tissue-of-origin, logistic regression was an informative system to explore the most accurate combination of markers, i.e., model 1 (*HUNK* and *RUNK1*), model 2 (*HUNK* only) and model 3 (*RUNX1* only). The main difference between double CpG-sites and each of the single CpG-site models was the detected imbalance between the sensitivity and specificity. While the 2-CpG-site model had a higher sensitivity than specificity (0.96 versus 0.82, respectively), the opposite was observed for the single-site models (0.78 versus 0.96 for model 2, and 0.81 versus 0.9 for model 3). Selection of the most accurate tissue prediction model was subsequently based on the additional metric of correct classification rate, with model 1 giving the best predictive performance of 88.59. Nevertheless, classifying these types of samples is complicated by the wide range of cellular proportions discussed above, as well as the admixed nature of some forensic specimens, e.g., cigarette butts. Therefore, for the second stage of the reported study, the generation of an age prediction model for oral cavity fluids which covered both saliva and buccal cells, was considered a better strategy than the development of different models for independent tissues. Even so, independent age prediction models for saliva and buccal cells were explored. Although the saliva-specific model showed the most accurate prediction (average MAE: ± 3.55), we decided to focus on the combined model since it will cover the maximum cell proportion variability in most forensic scenarios covering these samples.

To identify the most accurate age prediction model amongst nine saliva/buccal cell age correlated CpGs, different combinations of CpG sites were explored under multivariate quantile regression analysis testing up to eight different combined models (Table 4). Different age prediction models have been published based on different statistical tools, including linear regression [19,23–25,36,46,50], quadratic regression [28], machine learning [46] and quantile regression [20]. Although linear regression is the most commonly applied statistical analysis for age prediction, in this study quantile regression was selected, as its main advantage is the ability to provide age-specific prediction intervals, in addition to the predicted age.

From the CpG combinations tested, the selection of the most accurate age prediction model 7 was based on the best balance between error and the correct classification rate, with an MAE of ± 3.54, RMSE: 6.23 and % CP±PI: 76.08 %. Model 7 comprised CpG sites cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*, and discarding *FHL2* and *ASPA*. Their contribution to the tested models is insufficient to improve predictive performance. This was not unexpected given the low age-correlations displayed (0.198 and −0.332, respectively) plus high levels of tissue dispersion (SD= 0.145 and 0.162, respectively).

Previous age predictors targeting the oral cavity have been developed as tissue-independent models, obtaining prediction errors close to ± 5 years; including, Bocklandt et. al [9], Eipel et. al [22] and Schwender et. al [24] with reported MAEs of ± 5.2 (saliva), ± 4.3 years (buccal cells) and ± 5.11 years (buccal cells), respectively. Common to all three studies is the use of just three CpG sites compared to the seven of the present study, which could explain the higher prediction errors observed. Additional tissue-independent models presenting prediction errors similar to the present study such as Hong et. al [25], Jung et. al [23] and Wozniak et. al [50] with MAEs of ± 3.13 (saliva), ± 3.55 years (buccal cells) and ± 2.5 years (buccal cells), respectively, were based on 5–7 CpG sites. While the models presented by Hong et al. and Wozniak et al. are uniquely focused on saliva and buccal swab samples, respectively, the combined model developed by our study covers both tissues, being more reliable in forensic scenarios where a mixture of saliva and buccal cells is under study, such as cigarette butts. A similar strategy to the present study was developed by Jung et al. [23], building a 5-CpG tissue-combined age prediction model, including saliva, buccal swabs and blood samples. The prediction error obtained was MAE: ± 3.55, practically identical to the present study. Considering all these results and the fact that models of other tissues also systematically present errors close to ± 3 years, it is reasonable to conclude that the lowest error obtainable with current technologies has been reached. Independently of the tissues covered, the main improvement provided by the prediction model proposed in the present study in comparison to the previous ones is the underlying statistical method used – quantile regression – providing not only the predicted age but the age-specific prediction intervals as well. Since errors are usually narrower at younger samples rather than at older individuals, to provide a specific interval of ages could improve the accuracy of results.

Considering the models discussed above, it is evident that certain markers appear recurrently in multiple age predictors for saliva and buccal swabs, namely *ELOVL2*, *PDE4C*, *EDARADD* and *KLF14*. Genes *ELOVL2*, *PDE4C* and *EDARADD* are present in our model but with different CpG positions (except cg09809672 in *EDARADD*, shared with Schwender's [24] model). Comparing the markers in these four genes in the other studies shows that only the CpG of *PDE4C* is shared between Eipel's [22] and Schwender's [24] models. In *KLF14,* not used in our study, only cg14361627 is shared between Hong's [25] and Jung's [23] models. This CpG is in the list of 49 CpGs of the preselected markers (Supplementary Table S1) but did not meet the selection criteria for our model. Our marker selection was based on the GSE92767 dataset [25], the same dataset used for Hong's marker selection but different markers were selected by each study using the same dataset. Different approaches were used for marker selection by Hong, with linear regression

and stepwise regression used to identify markers with an R^2 greater than 0.65 and a difference between maximum and minimum β-scores greater than 0.1. This compares to our use of Spearman's correlation to select markers with a correlation greater than |0.8| and a difference between extreme age donors greater than |0.3|. The motivation to change the selection criteria for marker selection when assessing the GSE92767 dataset was based on the lack of normality found for 15 % of the residuals of the models (independent linear regression models for each CpG on the dataset). Therefore, a non-parametric method such as the Spearman coefficient was found to be more suitable for this analysis.

Regarding markers included in our prediction model, cg10501210 was reported as a marker related with aging in blood monocytes [51], showing a similar DNA methylation trend when analyzing saliva and buccal cells samples in our study. Although less evident than for *FHL2* and *ASPA*, the correlation with age and tissue dispersion detected for this marker ($r_s = -0.313$ and SD= 0.103) suggested exclusion form the final age prediction model. However, its removal from the final model has the greatest effect, as shown in the robustness analysis. The gene *LHFPL4* (LHFPL tetraspan subfamily member 4), is a member of the superfamily of tetraspan transmembrane protein encoding genes. Mutations in one LHFP-like gene result in deafness in humans and mice, and a second LHFP-like gene is fused to a high-mobility group gene in a translocation-associated lipoma. To the best of our knowledge, our study detected this marker to be correlated with age in saliva and buccal cells for the first time. The cg11084334 CpG analyzed in *LHFPL4* presented amongst the highest age correlation values ($r_s = 0.805$), as well as showing minimal dispersion between tissues (SD= 0.039). Correlation with age in blood has been observed in other CpG positions of *LHFPL4* (cg24866418 and cg12841266) [52]. The gene *ELOVL2* (ELOVL fatty acid elongase 2) has been widely reported as a key age correlated marker [12,53,54] and has been incorporated in most of the age prediction models developed so far. This marker has been reported to correlate with age in multiple forensic tissues such as blood [19,20,23,28,50], saliva [23], buccal cells [23,24,50], teeth [28] and bones [50]. More specifically, it is noteworthy that the cg16867657 CpG analyzed in our study has been reported in other studies to be correlated with age either in blood [19,20,28], buccal cells [24] or teeth [28]. Gene *PDE4C* (phosphodiesterase 4 C) had the strongest correlation with age in saliva and buccal cells was ($r_s = 0.806$), and has been published in age prediction models for different tissues including saliva, buccal cells and blood [18,20,22,28,37]. In gene *HOXC4* (homeobox C4), the cg18473521 CpG analyzed in this work has shown correlation with age in blood samples [55]. Gene *OTUD7A* (OTU deubiquitinase 7A), which encodes a protein acting on TNF receptor associated factor 6 (TRAF6) to control nuclear factor kappa B expression, is used for the first time in an age prediction model in our study. Although *OTUD7A* has previously shown correlation with age in blood [20] and saliva [25], it was not included in published any model. Finally, *EDARADD* has been reported to show age correlated CpG positions, with cg09809672 used in this study also reported in previous blood, saliva, buccal cell and bone models [9,20,24,28,50].

Our studies showed the age predictive performance of the saliva and buccal cell model was not improved by adding tissue-of-origin information. A similar analysis was performed by Eipel et. al for buccal swab samples [22]. In Eipel's study, combined age and cell-type prediction models reported age prediction errors with this model (training MAD ± 4.66; testing MAD ± 5.09) that improved on age correlated markers only (training MAD: ± 4.3; testing MAD: ± 7.03). This suggests that introducing the cellular composition as a co-variable has more effect than the tissue of origin. Therefore, assessment of the cellular proportions may be the most effective way to introduce tissue-of-origin information as a co-variable in an age prediction model – certainly for the buccal cavity.

Finally, considering that in forensic DNA analysis degraded and low-level DNA concentrations are commonly encountered, our evaluations of the robustness of the model with missing data and amounts of input DNA for bisulfite conversion were particularly relevant.

Similar predictive performance was obtained for all step-wise exclusions of markers with the exception of cg10501210 (MAE: ± 5.23, and %CP±PI=74.06 %). The absence of this CpG produced the greatest increase in error. However, it should be noted that if missing data are present, incorrect DNA methylation measurement could be also occurring at the detected methylated and unmethylated peaks. In this case, to run duplicates or even triplicates of the sample is recommended in order to double-check the methylation values obtained.

An important factor for forensic sensitivity of methylation tests is the bisulfite conversion step, representing an aggressive reduction of the input DNA. Since use of 100 ng is not common practice in casework, the serial dilutions that were evaluated up to 10 ng showed no standard deviations greater than 0.1. For 1 ng input, some markers showed deviation values above the established limit for 3 of 4 samples. Thus, it is a viable strategy to start with a minimum of 10 ng of genomic DNA. Very similar results have been obtained by Aliferi et.al [21] and Wózniak et.al [50], indicating analyses with less than 10 ng of DNA caused significant variations in DNA methylation values. When comparing these studies to data reported here, it is worth noting that different technologies have been used, massive parallel sequencing versus SNaPshot, suggesting the limitation is not the detection methodology, but the DNA degradation or loss during bisulfite conversion process itself, or the stochastic variability of the analyzed molecules.

Acknowledgements

This project was funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010) by a postdoctorate grant awarded to AFA. MVL is supported by the Ministerio de Educación, Cultura y Ciencia, Spain (PID2019-107876RB-I00).M.d.l.P. is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481D-2021-008). J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102770.

References

[1] A. Freire-Aradas, C. Phillips, M. Lareu, Forensic individual age estimation with DNA: from initial approaches to methylation tests, Forensic Sci. Rev. 29 (2) (2017) 121–144.
[2] W. Parson, Age estimation with DNA: From forensic DNA fingerprinting to forensic (Epi) genomics: a mini-review, Gerontology 64 (4) (2018) 326–332.
[3] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-kamysz, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int Genet 7 (1) (2013) 98–115.
[4] M. Marcińska, E. Pośpiech, S. Abidi, J.D. Andersen, M. van den Berge, Á. Carracedo, et al., Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness, PLoS One 10 (5) (2015), e0127852.
[5] A. Abbott, DNA clock may aid refugee age check, Nature 561 (2018) 15.
[6] R. Noroozi, S. Ghafouri-Fard, A. Pisarek, J. Rudnicka, M. Spólnicka, W. Branicki, et al., DNA methylation-based age clocks: From age prediction to age reversion, Ageing Res Rev. 68 (2021), 101314.
[7] Z.D. Smith, A. Meissner, DNA methylation: roles in mammalian development, Nat. Rev. Genet 14 (3) (2013) 204–220.
[8] V.K. Rakyan, T.A. Down, S. Maslau, T. Andrew, T.P. Yang, H. Beyan, et al., Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains, Genome Res 20 (4) (2010) 434–439.
[9] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sa, J.S. Sinsheimer, S. Horvath, et al., Epigenetic predictor of age, PLoS One 6 (6) (2011), e14821.

[10] J.T. Bell, P.-C. Tsai, T.-P. Yang, R. Pidsley, J. Nisbet, D. Glass, et al., Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population, PLoS Genet 8 (4) (2012), e1002629.
[11] H. Heyn, M. Esteller, DNA methylation profiling in the clinic: applications and challenges, Nat. Rev. Genet 13 (10) (2012) 679–692.
[12] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, Genome-wide methylation profiles reveal quantitative views of human aging rates, Mol. Cell 49 (2) (2013) 359–367.
[13] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (10) (2013) R115.
[14] A. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, PLoS One 8 (6) (2013), e67378.
[15] C.A. Reynolds, Q. Tan, E. Munoz, J. Jylhävä, J. Hjelmborg, L. Christiansen, et al., A decade of epigenetic change in aging twins: genetic and environmental contributions to longitudinal DNA methylation, Aging Cell 19 (8) (2020), e13197.
[16] Y. Wang, R. Karlsson, E. Lampa, Q. Zhang, Å.K. Hedman, M. Almgren, Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins, Epigenetics 13 (9) (2018) 975–987.
[17] L.D. Moore, T. Le, G. Fan, DNA methylation and its basic function, Neuropsychopharmacology 38 (1) (2013) 23–38.
[18] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al., Aging of blood can be tracked by DNA methylation changes at just three CpG sites, Genome Biol. 15 (2) (2014) R24.
[19] R. Zbieć-Piekarska, M. Sólnicka, T. Kupiec, A. Parys-proszek, Z. Makowska, A. Paleczka, et al., Development of a forensically useful age prediction method based on DNA methylation analysis, Forensic Sci. Int Genet 17 (2015) 173–179.
[20] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares De Cal, 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 24 (2016) 65–74.
[21] 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.
[22] M. Eipel, F. Mayer, T. Arent, M.R.P. Ferreira, C. Birkhofer, U. Gerstenmaier, et al., Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures, Aging 8 (5) (2016) 1034–1048.
[23] S.-E. Jung, S. Min, S. Rom, E. Hee, K. Shin, H. Young, 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.
[24] K. Schwender, O. Holländer, S. Klopfleisch, M. Eveslage, M.F. Danzer, H. Pfeiffer, et al., Development of two age estimation models for buccal swab samples based on 3 CpG sites analyzed with pyrosequencing and minisequencing, Forensic Sci. Int Genet 53 (2021), 102521.
[25] 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.
[26] W.J. Lee, C.M. Choung, Y.J. Jung, H.Y. Lee, S.-K. Lim, A validation study of DNA methylation-based age prediction using semen in forensic casework samples, Leg. Med 31 (2018) 74–77.
[27] T.G. Jenkins, K.I. Aston, B. Cairns, A. Smith, D.T. Carrell, Paternal germ line aging: DNA methylation age prediction from human sperm, BMC Genom. 19 (1) (2018) 763.
[28] B. Bekaert, A. Kamalandua, S.C. Zapico, W.Van De Voorde, B. Bekaert, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, Epigenetics 10 (10) (2015) 922–930.
[29] H.Y. Lee, S.R. Hong, J.E. Lee, I.K. Hwang, N.Y. Kim, J.M. Lee, et al., Epigenetic age signatures in bones, Forensic Sci. Int Genet 46 (2020), 102261.
[30] L.E. Reinius, N. Acevedo, M. Joerink, G. Pershagen, S.-E. Dahlén, D. Greco, et al., Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility, PLoS One 7 (7) (2012), e41361.
[31] C. Theda, S.H. Hwang, A. Czajko, Y.J. Loke, P. Leong, J.M. Craig, Quantitation of the cellular content of saliva and buccal swab samples, Sci. Rep. 8 (1) (2018) 6944.
[32] S. Horvath, J. Oshima, G.M. Martin, A.T. Lu, A. Quach, S. Felton, et al., Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies, Aging (Albany NY) 10 (7) (2018) 1758–1775.
[33] S. Köchl, H. Niederstätter, W. Parson, DNA extraction and quantitation of forensic samples using the phenol-chloroform method and real-time PCR, Methods Mol. Biol. 297 (2005) 13–30.
[34] R.C. Slieker, S.D. Bos, J.J. Goeman, J.V.M.G. Bovée, R.P. Talens, R. Breggen, Van Der, et al., Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array, Epigenetics Chromatin 6 (1) (2013) 26.
[35] M.J. Jones, P. Farré, L.M. Mcewen, J.L. Macisaac, K. Watt, S.M. Neumann, et al., Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in down syndrome, BMC Med Genom. 6 (2013) 58.
[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.
[37] J.J. Marqueta-Gracia, M. Álvarez-Álvarez, M. Baeta, L. Palencia-Madrid, E. Prieto-Fernández, R.J. Ordoñana, et al., Genetics differentially methylated CpG regions analyzed by PCR-high resolution melting for monozygotic twin pair discrimination, Forensic Sci. Int Genet 37 (2018) e1–e5.
[38] Y. Hamano, S. Manabe, C. Morimoto, S. Fujimoto, K. Tamaki, Forensic age prediction for saliva samples using methylation-sensitive high resolution melting: exploratory application for cigarette butts, Sci. Rep. 7 (5) (2017) 10444.
[39] F.M. You, N. Huo, Y.Q. Gu, M. Luo, Y. Ma, D. Hane, et al., BatchPrimer3: a high throughput web application for PCR and sequencing primer design, BMC Bioinforma. 9 (2008) 253.
[40] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinforma. 12 (2011) 77.
[41] Koenker R., Portnoy S., Ng P., Zeileis A., Grosjean P., Ripley B. Package quantreg: Quantile Regression. 2015.
[42] Alfons A. Package cvTools: Cross-validation tools for regression models. 2015.
[43] Wickham H., Chang W. Package ggplot2: An implementation of the grammar of graphics. 2015.
[44] Team R Core. R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,, Vienna, Austria, 2020 (Available from), ⟨https://www.r-project.org/⟩.
[45] A. Vidaki, B. Daniel, D.S. Court, Forensic DNA methylation profiling — potential opportunities and challenges, Forensic Sci. Int Genet 7 (5) (2013) 499–507.
[46] S.R. Hong, K. Shin, S. 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.
[47] H. Alghanim, K. Balamurugan, B. Mccord, Development of DNA methylation markers for sperm, saliva and blood indentification using pyrosequencing and qPCR/HRM, Anal. Biochem 611 (2020), 113933.
[48] C. Thiede, G. Prange-Krex, J. Freiberg-Richter, M. Bornhäuser, G. Ehninger, Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants, Bone Marrow Transpl. 25 (5) (2000) 575–577.
[49] W.P. Accomando, J.K. Wiencke, E.A. Houseman, H.H. Nelson, K.T. Kelsey, Quantitative reconstruction of leukocyte subsets using DNA methylation, Genome Biol. 15 (3) (2014) R50.
[50] A. Wózniak, A. Heidegger, D. Piniewska-Róg, E. Pośpiech, C. Xavier, A. Pisarek, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones, Aging 13 (5) (2021) 6459–6484.
[51] L. Tserel, M. Limbach, M. Saare, K. Kisand, A. Metspalu, L. Milani, et al., CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing, Immun. Ageing 11 (1) (2014) 1.
[52] H. Alsaleh, P.R. Haddrill, Identifying blood-specific age-related DNA methylation markers on the Illumina methylationEPIC BeadChip, Forensic Sci. Int 303 (2019), 109944.
[53] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, et al., Methylation of ELOVL2 gene as a new epigenetic marker of age Aging Cell, Aging 11 (6) (2012) 1132–1134.
[54] H. Heyn, N. Li, H.J. Ferreira, S. Moran, D.G. Pisano, A. Gomez, et al., Distinct DNA methylomes of newborns and centenarians, Proc. Natl. Acad. Sci. USA 109 (26) (2012) 10522–10527.
[55] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-hoekstra, M.C.H.Van Der Zwalm, P. 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.

Artículo 3: Development of an epigenetic age predictor for costal cartilage with a simultaneous somatic tissue differentiation system

A. Freire-Aradas, M. Tomsia, D. Piniewska-Róg, A. Ambroa-Conde, M. A. Casares de Cal, A. Pisarek, A. Gómez-Tato, J. Álvarez-Dios, E. Pośpiech, W. Parson, M. Kayser, C. Phillips, W. Branicki.

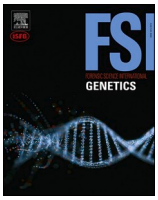
Forensic Science International: Genetics, volumen 67, página 102936, 2023 DOI: 10.1016/j.fsigen.2023.102936

Forensic Science International: Genetics 67 (2023) 102936

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



Development of an epigenetic age predictor for costal cartilage with a simultaneous somatic tissue differentiation system

A. Freire-Aradas [a,*], M. Tomsia [b], D. Piniewska-Róg [c], A. Ambroa-Conde [a], MA Casares de Cal [d], A. Pisarek [e], A. Gómez-Tato [d], J. Álvarez-Dios [f], E. Pośpiech [g,h], W. Parson [i,j], M. Kayser [k], C. Phillips [a], W. Branicki [e,l,**]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] Department of Forensic Medicine and Forensic Toxicology, Medical University of Silesia, Katowice, Poland
[c] Department of Forensic Medicine, Jagiellonian University Medical College, Kraków, Poland
[d] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain
[e] Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland
[f] Faculty of Mathematics, University of Santiago de Compostela, Spain
[g] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[h] Department of Forensic Genetics, Pomeranian Medical University in Szczecin, Poland
[i] Institute of Legal Medicine, Medical University of Innsbruck, Austria
[j] Forensic Science Program, Pennsylvania State University, PA, USA
[k] Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[l] Institute of Forensic Research, Kraków, Poland

ARTICLE INFO

Keywords:
DNA methylation
age estimation
costal cartilage
tissue identification
bone
blood
buccal cells

ABSTRACT

Age prediction from DNA has been a topic of interest in recent years due to the promising results obtained when using epigenetic markers. Since DNA methylation gradually changes across the individual's lifetime, prediction models have been developed accordingly for age estimation. The tissue-dependence for this biomarker usually necessitates the development of tissue-specific age prediction models, in this way, multiple models for age inference have been constructed for the most commonly encountered forensic tissues (blood, oral mucosa, semen). The analysis of skeletal remains has also been attempted and prediction models for bone have now been reported. Recently, the VISAGE Enhanced Tool was developed for the simultaneous DNA methylation analysis of 8 age-correlated loci using targeted high-throughput sequencing. It has been shown that this method is compatible with epigenetic age estimation models for blood, buccal cells, and bone. Since when dealing with decomposed cadavers or postmortem samples, cartilage samples are also an important biological source, an age prediction model for cartilage has been generated in the present study based on methylation data collected using the VISAGE Enhanced Tool. In this way, we have developed a forensic cartilage age prediction model using a training set composed of 109 samples (19–74 age range) based on DNA methylation levels from three CpGs in *FHL2*, *TRIM59* and *KLF14*, using multivariate quantile regression which provides a mean absolute error (MAE) of ± 4.41 years. An independent testing set composed of 72 samples (19–75 age range) was also analyzed and provided an MAE of ± 4.26 years. In addition, we demonstrate that the 8 VISAGE markers, comprising *EDARADD*, *TRIM59*, *ELOVL2*, *MIR29B2CHG*, *PDE4C*, *ASPA*, *FHL2* and *KLF14*, can be used as tissue prediction markers which provide reliable blood, buccal cells, bone, and cartilage differentiation using a developed multinomial logistic regression model. A training set composed of 392 samples (n = 87 blood, n = 86 buccal cells, n = 110 bone and n = 109 cartilage) was used for building the model (correct classifications: 98.72%, sensitivity: 0.988, specificity: 0.996) and validation was performed using a testing set composed of 192 samples (n = 38 blood, n = 36 buccal cells, n = 46 bone and n = 72 cartilage) showing similar predictive success to the training set (correct classifications: 97.4%, sensitivity: 0.968, specificity: 0.991). By developing both a new cartilage age model and a tissue differentiation model, our study significantly expands the use of the VISAGE Enhanced Tool while

* Corresponding author.
** Corresponding author at: Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland.
E-mail addresses: ana.freire@usc.es (A. Freire-Aradas), wojciech.branicki@uj.edu.pl (W. Branicki).

https://doi.org/10.1016/j.fsigen.2023.102936
Received 14 April 2023; Received in revised form 13 September 2023; Accepted 27 September 2023
Available online 29 September 2023

increasing the amount of DNA methylation-based information obtained from a single sample and a single forensic laboratory analysis. Both models have been placed in the open-access Snipper forensic classification website.

1. Introduction

The identification of cadavers and human remains recovered from mass disaster events [1,2], exhumation [3–5], or missing persons identification [6] are among the many tasks performed by forensic geneticists. New technologies using biometric data, including machine learning methods, increasingly provide support for conventional human identification methods [7], but still the FBI and Interpol annually disclose their databases of unidentified remains [8,9]. According to the FBI's National Crime Information Center report for 2022, this database listed 546,568 missing person records. The 2022 records included 581 (64%) unidentified human cadavers, 5 (1%) unidentified mass disaster victims, and 316 (35%) living persons who could not ascertain their identity [8].

Costal cartilage tissue has been successfully used in cases of individual identification of corpses at an advanced stage of decomposition or with almost complete skeletonization and can be an alternative to soft tissue that has degraded DNA [10–12]. In fact, costal cartilage is one of the last tissues available for sampling before complete skeletonization occurs in human remains [13]. It is worth noting that cartilage tissue is a practical alternative forensic material to bone and teeth due to a much faster and cheaper DNA extraction procedure [10,14,15]. The advantages of using costal cartilage in the process of human identification has been shown in some cases because it may bypass the problem of chimerism [16,17]. In addition, cartilage tissue is increasingly chosen for other forensic analyses including forensic toxicology [18–21].

The hyaline tissue of the costal cartilage is the thickest cartilage in the human body. It connects the ribs to the sternum [22] has an abundant extracellular matrix that protects its DNA from environmental factors, and, amongst other material, consists of proteoglycans, and collagen I and II. As age increases, the quantity of type I collagen gradually decreases compared to type II collagen [23]. Pfeiffer et al. [24] demonstrated that the D/L aspartic acid ratios in the insoluble collagen fraction correlate with age ($r = 0.97$) and noted the 95% confidence interval around an estimated individual age was approximately 14 years. Other studies have shown that the extent of costal cartilage calcification (the measurement of calcification foci) might correlate with age [25,26]. Since the layer of hyaline tissue on long bone articular surfaces is much thinner, and thus more susceptible to degradation [27], the correlation between DNA methylation age and chronological age was found to be weaker ($r = 0.79$) and a median absolute difference of four years was estimated [28]. Age estimation can help identify decaying human remains by limiting the search criteria for a missing person but the forensic utility of a predictive method depends primarily on its accuracy [29]. It was found that epigenetic changes are essential in the aging process, and consequently DNA methylation (mDNA) has proven to be a very accurate measure of an individual's age. However, tissue specificity of DNA methylation patterns complicates the development of a universal epigenetic age predictor in forensics. The epigenetic age prediction has been explored more extensively for body fluids including blood and saliva which are often examined in criminal cases [30–32]. Recently, the VISAGE Consortium introduced the VISAGE Enhanced Tool for age prediction from somatic tissues based on eight most relevant methylation markers, covering 44 individual CpG sites. An assay based on target massively parallel sequencing (MPS) technology using the MiSeq system (Illumina, San Diego, CA, USA) was developed for simultaneous analysis of these markers enabling the collection of methylation data to predict age using specific algorithms for blood, buccal cells, as well as bone [32]. Age prediction models for blood, buccal cells and bone based on tissue-specific model training and testing datasets have been established providing prediction accuracies with mean absolute error of ± 3.2 to ± 3.7 years for the tissue-specific models.

A method of age prediction based on costal cartilage can be potentially useful in cases of identification of cadavers at an advanced stage of decomposition, when soft tissues or biological fluids are too degraded [33]. An additional reason for developing a predictive model for rib cartilage was that the models developed in our previous studies for blood, oral mucosa, and bone did not correctly predict the age of cartilage. Therefore, the present study aimed to develop a cartilage age prediction model based on three CpG sites analyzed with the VISAGE Enhanced Tool for age prediction from somatic tissues [32] using multivariate quantile regression as the predictive framework. A total of 109 costal cartilage samples ranging from 19 to 74 years were used to train the model. K-fold cross-validation was used for validation purposes, as well as an independent testing set composed of 72 samples from 19 to 75 years. Additionally, the VISAGE Enhanced Tool was adapted for tissue identification purposes covering the cell types targeted for age inference. Thus, a tissue prediction model for differentiating blood, buccal cells, bone, and cartilage was created using multinomial logistic regression based on a training set composed of 392 samples. We used k-fold cross-validation for validation, as well as a testing set composed of 192 samples, both training and test sets contained the four tissues of interest.

2. Material and Methods

2.1. Samples, DNA extraction, and quantification

The study was approved by the Bioethics Committee of the Jagiellonian University in Kraków, Poland (KBET/122.6120.86.2017). A total of 181 costal cartilage samples were collected during medicolegal autopsies at the Department of Forensic Medicine, Jagiellonian University Medical College in Krakow, Poland. The study group consisted of 145 males and 36 females in the age range of 19–75 years (mean 44.3 ± 14.6) and with a time from death to autopsy ranging from 1 to 5 days. The costal cartilage portions (5 × 6 cm) were collected from the cadavers' rib arches and stored at − 80 °C until further processing. Before starting the genetic analysis, the sampled material was cleaned by removing its external surface and any contamination, using a sterile scalpel, and then fragmented into small cubes. The fragmented tissue was placed in 1.5 ml Eppendorf tubes and incubated in an extraction mixture (Sherlock AX kit, A&A Biotechnology, Poland) at 50 °C in a Thermo–Shaker TS-100 C (Biosan) at 500 rpm for 24 h. Total DNA was extracted using a silica-based method with the Sherlock AX kit, according to the manufacturer's protocol. In addition, a total of 125 blood samples comprising 62 males and 63 females in the age range of 19–75 years old, (mean 48.9 ± 17.5); 122 buccal swabs comprising 62 males and 60 females in the age range of 19–80 years (mean 49.25 ± 17.9) and 156 bone samples comprising 125 males and 31 females in the age range of 19–75 years (mean 45.9 ± 14.4), were collected in the VISAGE project as described in Woźniak et al. [32] and used in this study. The quality and quantity of DNA isolates were measured using a NanoDrop 8000 UV-Vis Spectrophotometer and Qubit dsDNA HS Assay Kit on a Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), following manufacturer's guidelines.

2.2. Bisulfite conversion, multiplex PCR, and massively parallel sequencing

Bisulfite conversion (BC) was performed with 500 ng of DNA using the EZ DNA Methylation-Direct Kit (Zymo Research, Irvine, CA, USA)

and then eluted in 25 μL.

DNA methylation levels were quantified using the VISAGE Enhanced Tool age prediction system [32]. The VISAGE MPS assay is based on PCR enrichment of targeted regions from bisulfite-converted DNA [34] and analyses of 44 CpG sites in eight age informative markers, namely *KLF14*, *TRIM59*, *MIR29B2CHG, FHL2*, *ELOVL2*, *EDARADD, PDE4C* and *ASPA*. In brief, 5 μL of the bisulfite-converted DNA samples were amplified in one multiplex PCR assay and libraries prepared using the KAPA Hyper Prep Kit and KAPA Unique-Dual Indexed Adapters (Roche, Basel, Switzerland). All samples were sequenced on the MiSeq instrument using the MiSeq Reagent Kit v3 (600 cycles) (Verogen, San Diego, CA, USA) with 2 × 150 cycles. Control and test samples were sequenced together in one run if the used indexes allowed multiplexing to avoid batch effects. Pooled libraries were diluted to 12pM and sequenced with a 10% PhiX spike.

Raw data in FASTQ files were aligned against a custom reference genome using a bwa-meth method as described in detail in Woźniak et al. [32]. The number of reads at the target positions were extracted using bam-readcount with a minimum mapping quality and minimum base quality set to 30 (https://github.com/genome/bam-readcount); DNA methylation levels were calculated as the ratio of C reads to the sum of C and T reads and expressed as a percentage (C reads/(C reads + T reads) * 100).

2.3. Statistical analyses

All calculations were performed using R software v.4.2.2 applying R scripts developed in-house. Spearman correlation (r_s) was used to assess correlations between DNA methylation levels and chronological age. Multivariate quantile regression with quantiles 0.1 and 0.9 was used to build the age prediction model with the *quantreg* R package [35]. The corresponding predictive accuracy was measured with the following performance metrics: mean absolute error (MAE) and root-mean-square error (RMSE). Although when working with quantiles, the MAE can be represented by the median instead of the mean, the mean was used in the present study for comparative purposes with additional models. Predicted versus chronological age was plotted using the *ggplot2* R package [36]. The final online cartilage age prediction model developed in our study has been placed in the open access Snipper forensic classification website and is freely available at: http://mathgene.usc.es/cgi-bin/snps3/age_tools/processmethylation-cartilage.cgi. Multinomial logistic regression was used for the development of the tissue prediction model using the *nnet* R package [37]. The corresponding predictive accuracy was measured with the following performance metrics: percent of correct classifications (%CC), sensitivity and specificity. Principal components were plotted using the *factoextra* R package [38]. Validation of the prediction models was performed using k-fold cross-validation (k = 10). The k-fold cross-validation system randomly cleaves input data into k fragments of similar sample size. Random cleavage of the input data was made using the *cvTools* R package [39]. Every k time that the model was assessed, a k cluster was retained as the test set with the remaining clusters used as the training set, maintaining proportions of 10% and 90% of the input data for test and training sets respectively, per run. The final online tissue prediction model developed in our study has been placed in the open access Snipper forensic classification website and is freely available at: http://mathgene.usc.es/cgi-bin/snps3/tissue_tools/processmethylation-tissue.cgi.

3. Results

3.1. CpG selection for age estimation in cartilage samples

A total of 44 CpGs located in eight genes (*KLF14*, *TRIM59*, *MIR29B2CHG, FHL2*, *ELOVL2*, *EDARADD, PDE4C* and *ASPA*) were analyzed in 181 costal cartilage samples. Specific information for the CpGs analyzed is outlined in Supplementary Table S1 and dispersion diagrams are shown in Supplementary Fig. S1. The corresponding DNA methylation data can be found in Supplementary Table S2. Three genes were discarded for the subsequent age estimation analyses due to a lack of, or weak correlation with age: *MIR29B2CHG*, *EDARADD* and *ASPA* (mean r_s: 0.04, −0.113 and 0.166, respectively). Patterns of hypermethylation were shown for the remaining five genes. The gene that had the highest correlation with age was *FHL2* for the CpG C4 (r_s: 0.931), followed by *TRIM59* for C7 (r_s: 0.906), *KLF14* for C3 (r_s: 0.888), *PDE4C* for C6 (r_s: 0.802) and lastly *ELOVL2* for C9 (r_s: 0.794). These eight CpG sites were initially selected for building the cartilage age prediction model.

3.2. Development of an age prediction model for cartilage samples

From a total of 181 costal cartilage samples, a subset of 109 (19–74 years old) was used for training the cartilage age prediction model. Multivariate quantile regression was tested with one-step increases in the total number of CpG sites included. The CpG sites were retained in the successive models following a decreasing order of correlation with age, and the addition of CpG sites stopped when no additional improvement was apported to the model. Table 1 shows the corresponding cross-validation performance metrics for the training set for the four assessed models (M1, M2, M3 and M4) In this way, the CpG with the highest Spearman correlation was assessed in the first model M1 (*FHL2*_C4), providing an MAE of ± 4.61 years and RMSE of 5.81. For the addition of a second CpG, *TRIM59*_C7 was analyzed (M2). Results obtained from M2 (*FHL2*_C4 and *TRIM59*_C7), provided an MAE of ± 4.47 years and RMSE of 5.76. The third most age-correlated CpG of *KLF14*_C3 was included in M3 (*FHL2*_C4, *TRIM59*_C7 and *KLF14*_C3) showing an MAE of ± 4.41 years and RMSE of 5.52. The fourth most age-correlated CpG was *PDE4C*_C6 included in M4 (*FHL2*_C4, *TRIM59*_C7, *KLF14*_C3 and *PDE4C*_C6). This model provided an MAE of ± 4.41 years and RMSE of 5.53. Since M4 do not improve the previous model M3, one-step increases were stopped at this point.

From these analyses, the optimum age prediction model was identified to be M3 (*FHL2*_C4, *TRIM59*_C7 and *KLF14*_C3) since this model provided the most accurate predictions. Therefore, M3 was selected as the final age prediction model for cartilage samples. To validate the final model, an independent subset of 72 samples (19–75 years old) was analyzed and provided an MAE of ± 4.26 years and RMSE of 5.39. Figs. 1A and 1B show the predicted age versus the chronological age for the training and test sets assessed with the M3 model. Continuous grey and black lines represent perfect and fitted correlations, respectively, while discontinuous black lines represent the age-specific prediction intervals. We detected correlation of the predicted with the chronological age with an R^2 value of 0.863 and 0.864 for the training and test sets, respectively. However, a slight overestimation of age in young

Table 1
Predictive performance metrics for the training (cross-validation) and test sets in cartilage for four age prediction models tested using quantile regression. MAE: Mean Absolute Error, RMSE (Root-Mean-Square Error). Bold indicates the selected model.

	Model	CpG nº	CpG_ID	MAE	RMSE
Training age cartilage (n = 109)	M1	1	*FHL2*_C4	± 4.61	5.81
	M2	2	*FHL2*_C4 *TRIM59*_C7	± 4.47	5.76
	M3	**3**	***FHL2*_C4** ***TRIM59*_C7** ***KLF14*_C3**	**± 4.41**	**5.52**
	M4	4	*FHL2*_C4 *TRIM59*_C7 *KLF14*_C3 *PDE4C*_C6	± 4.41	5.53
Test age cartilage (n = 72)	**M3**	**3**	***FHL2*_C4** ***TRIM59*_C7** ***KLF14*_C3**	**± 4.26**	**5.39**

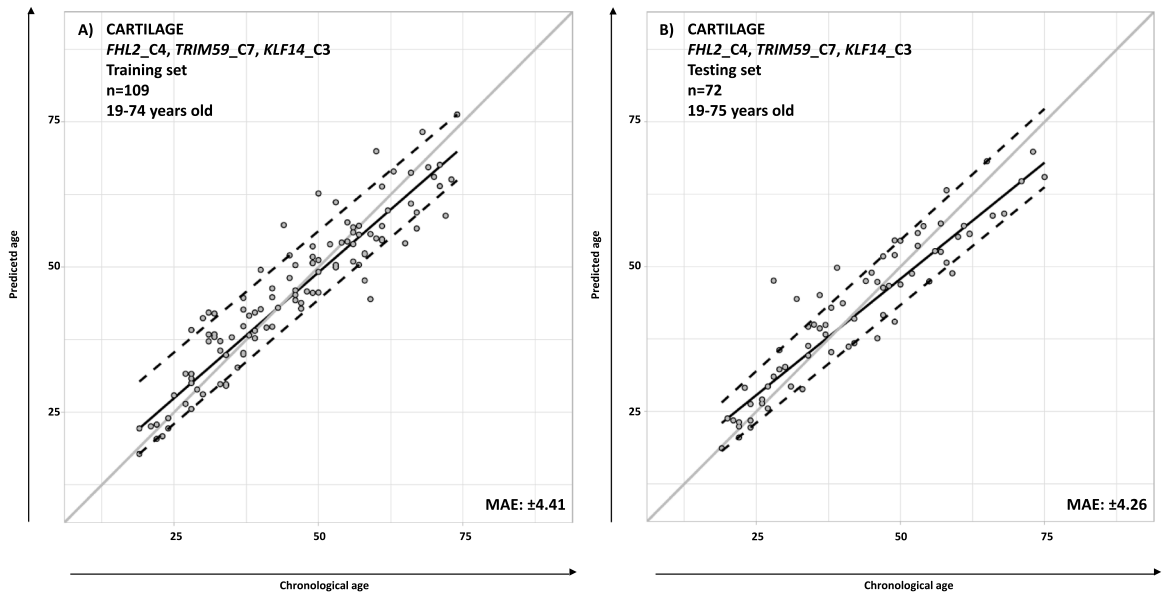


Fig. 1. Plot of predicted age against chronological age for cartilage in: **A)** the training set (n = 109, 19–74 years old); and **B)** the test set (n = 72, 19–75 years). Predicted age was inferred from a multivariate quantile regression model based on three CpG sites (*FHL2*_C4, *TRIM59*_C7 and *KLF14*_C3). The black diagonal line represents the 0.5 quantile regression line between predicted age and chronological age and the discontinuous black lines, the corresponding 0.1 and 0.9 quantile regression limits (prediction intervals). The grey line is the diagonal line representing perfect correlation.

individuals as well as a moderate underestimation in older samples is observed in the test set.

The following equation describes the multivariate quantile regression model M3:

Predicted age in years = −0.79 + (0.65 x *FHL2*_C4) + (0.04 x *TRIM59*_C7) + (2.05 x *KLF14*_C3).

Prediction intervals were estimated using the following equations:

Minimum Prediction (MinPred, q 0.1) = 1.16 + (0.33 x *FHL2*_C4) + (0.47 x *TRIM59*_C7) + (1.93 x *KLF14*_C3).

Maximum Prediction (MaxPred, q 0.9) = −2.55 + (0.79 x *FHL2*_C4) - (0.06 x *TRIM59*_C7) + (2.28 x *KLF14*_C3).

The final cartilage age prediction model M3 is freely available from the open-access Snipper forensic classification website described in Material and Methods.

3.3. CpG selection for the inference of the tissue of origin

In addition to the 181 cartilage samples, the 44 CpGs located in *KLF14*, *TRIM59*, *MIR29B2CHG, FHL2*, *ELOVL2*, *EDARADD, PDE4C* and *ASPA* were also analyzed in 125 blood samples, 122 buccal swabs and 156 bone samples, using DNA methylation data gathered in Woźniak et al. The corresponding DNA methylation data can be found in Supplementary Table S3. The corresponding dispersion diagrams are shown in Supplementary Fig. S2 and the mean DNA methylation values per tissue for all 44 CpG sites in Supplementary Table S4. The DNA methylation levels for two of the genes previously discarded for age estimation – *MIR29B2CHG* and *EDARADD* – showed a marked difference between cartilage samples and the remaining somatic tissues. The three CpG sites analyzed for *MIR29B2CHG* (C1, C2 and C3) showed continuous patterns of methylation of about ~10–60% (mean: 33.81%) in cartilage samples, while higher methylation levels were found for blood (mean: 75.01%), buccal cells (mean: 78.53%) and bone (mean: 85.17%). The opposite pattern was detected in *EDARADD* (C1 and C2). Continuous hypermethylation was observed in cartilage samples (mean: 83.47%), whereas the additional somatic tissues had lower levels of methylation (mean: 38.75%). The *ASPA* gene previously discarded for age estimation, was informative for blood samples with the highest differences in DNA methylation values (mean: 72.16%) compared to other somatic tissues (mean: 32.51%). In genes *KLF14*, *TRIM59*, *FHL2*, *ELOVL2* and *PDE4C*, patterns of tissue differentiation were evident in most of the CpGs in these genes. In the case of *KLF14*, C1 gave the highest differences in DNA methylation levels between bone (mean: 12.9%) and other tissues (mean: 6.5%). In *TRIM59*, although some overlapping values were detected between blood and buccal cells for young individuals, methylation differences were observed among the four tissues – blood, buccal cells, bone, and cartilage – for C1 (mean: 25.73%, 34.82%, 14% and 7.3%, respectively), C2 (mean: 20.49%, 28.25%, 9.93% and 5.62%, respectively), C3 (mean: 29.92%, 46.35%, 17.32% and 12.74%, respectively), C4 (mean: 47.22%, 61.37%, 29.07% and 18.85%, respectively) and C5 (mean: 45.99%, 62.13%, 26.58% and 19.27%, respectively); while C3, C6, C7 and C8 had overlapping values between bone and cartilage. In *FHL2*, the highest differences in DNA methylation levels were detected between blood samples and other tissues (difference in DNA methylation levels for blood versus the remaining tissues: 32.92%, 34.09%, 33.04%, 34.11%, 30.19%, 29.54%, 16.32%, 13.65%, 28.36% and 8.97%, from C1 to C10, respectively). The differentiation of buccal cells was also achieved with *FHL2*, but partially, since a portion of samples had marked levels of hypomethylation, but the remaining proportion overlapped with bone and cartilage. *ELOVL2* had methylation differences among the four tissues for all the C1 to C9 analyzed CpGs (mean: 49.53%, 63.33%, 28.15% and 16.85% for blood, buccal cells, bone, and cartilage, respectively). In *PDE4C*, simultaneous differentiation of the four tissues was mainly shown by C3 (mean: 33.36%, 25.1%, 11.91% and 7.01%, for blood, buccal cells, bone, and cartilage, respectively).

To select the most informative CpGs, each marker was assessed using multinomial logistic regression. Supplementary Table S5 shows the corresponding mean percentage of correct classifications for all the 44 CpGs analyzed in the eight genes. The following CpGs gave the highest rate of correct classifications per gene and were selected for further analyses: *EDARADD*_C1 (76.37%), *TRIM59*_C1 (76.54%), *ELOVL2*_C6 (72.09%), *MIR29B2CHG*_C3 (71.4%), *PDE4C*_C3 (70.03%), *ASPA*_C1 (63.53%), *FHL2*_C10 (61.3%) and *KLF14*_C1 (51.2%). These eight CpG sites were then selected to build a tissue prediction model.

3.4. A multinomial logistic prediction model for somatic tissue differentiation

Based on the DNA methylation differences amongst different tissues outlined above, a tissue prediction model was developed using logistic regression. In order to build the model, a training set of 392 samples was assessed (n = 87 blood samples, n = 86 buccal swabs, n = 110 bone and n = 109 cartilage samples). Using this training set, a total of eight statistical models based on multinomial logistic regression were tested with one-step increases in the number of CpG sites included (M1, M2, M3, M4, M5, M6 and M7 and M8). Table 2 shows the corresponding cross-validation performance metrics for the training set for the eight evaluated models. The CpG sites were serially included in the models following a decreasing order of correct classification rate, as described above. Using this approach, the CpG site providing the highest level of correct classifications was evaluated in the first model M1 (*EDARADD*_C1) correctly classifying 77.3% of the samples and showing a sensitivity of 0.76 and a specificity of 0.924. Each time an additional CpG site was incorporated into the model, including *TRIM59*_C1 in M2, *ELOVL2*_C6 in M3, *MIR29B2CHG*_C3 in M4, *PDE4C*_C3 in M5, *ASPA*_C1 in M6, *FHL2*_C10 in M7 and *KLF14*_C1 in M8; the rate of correct classifications increased to reach a maximum of 98.72% for M8 (all the eight CpGs sites included), as well as 0.988 for sensitivity and 0.996 for specificity, so M8 was selected as the final tissue prediction model. Listed per tissue, the following results were obtained for blood (98.85% correct classification, 0.989 sensitivity, 0.993 specificity), for buccal swabs (%CC: 100%, sensitivity: 1 specificity: 0.997), for bone (%CC: 96.36%, sensitivity: 0.964 specificity: 1) and for cartilage samples (%CC: 100%, sensitivity: 1 specificity: 0.993). Principal Component Analysis (PCA) results based on the eight CpGs (*EDARADD*_C1, *TRIM59*_C1, *ELOVL2*_C6, *MIR29B2CHG*_C3, *PDE4C*_C3, *ASPA*_C1, *FHL2*_C10 and *KLF14*_C1) are plotted in Fig. 2A. Blood samples, buccal swabs, bone, and cartilage samples are represented as red, orange, violet and green datapoints, respectively. The first Principal Component (PC1) explains 49.7% of tissue separation, and the second, PC2, 27.9%. To validate the tissue prediction model, a test set of 192 samples (n = 38 blood samples, n = 36 buccal swabs, n = 46 bone and n = 72 cartilage samples) was assessed and gave 97.4% correct classifications, a sensitivity of 0.968 and specificity of 0.991. Listed per tissue these gave the following values for blood (%CC: 97.3%, sensitivity: 0.973 specificity: 0.987), for buccal swabs (%CC: 89.74%, sensitivity: 0.897 specificity: 0.993), for bone (%CC: 100%, sensitivity: 1 specificity: 0.993) and for cartilage samples (%CC: 100%, sensitivity: 1 specificity: 0.992). The corresponding PCA is shown in Fig. 2B. The main difference in comparison to the training set was the decrease in the classification rate for buccal swabs (from 100% to 89.74%). This is due to four misclassifications (two samples predicted as blood, one sample as bone and one as cartilage). The remaining tissues gave similar classification rates between the training and testing sets. The final tissue prediction model is freely available from the open-access Snipper forensic classification website described in Material and Methods.

Table 2
Predictive performance metrics for the training (cross-validation) and testing set for the somatic tissue prediction model developed using multinomial logistic regression (blood vs buccal cells vs bone vs cartilage samples). $\%CC_{mean}$ (mean of the percent of correct classifications). Bold indicates the selected model.

	Model	CpG nº	CpG_ID	$\%CC_{mean}$	$Sensitivity_{mean}$	$Specificity_{mean}$
Training tissue (n = 392)	M1	1	*EDARADD*_C1	77.3%	0.76	0.924
	M2	2	*EDARADD*_C1 *TRIM59*_C1	86.97%	0.862	0.957
	M3	3	*EDARADD*_C1 *TRIM59*_C1 *ELOVL2*_C6	89.28%	0.884	0.965
	M4	4	*EDARADD*_C1 *TRIM59*_C1 *ELOVL2*_C6 *MIR29B2CHG*_C3	91.84%	0.913	0.973
	M5	5	*EDARADD*_C1 *TRIM59*_C1 *ELOVL2*_C6 *MIR29B2CHG*_C3 *PDE4C*_C3	96.18%	0.958	0.987
	M6	6	*EDARADD*_C1 *TRIM59*_C1 *ELOVL2*_C6 *MIR29B2CHG*_C3 *PDE4C*_C3 *ASPA*_C1	96.68%	0.965	0.989
	M7	7	*EDARADD*_C1 *TRIM59*_C1 *ELOVL2*_C6 *MIR29B2CHG*_C3 *PDE4C*_C3 *ASPA*_C1 *FHL2*_C10	98.21%	0.98	0.994
	M8	**8**	***EDARADD*_C1** ***TRIM59*_C1** ***ELOVL2*_C6** ***MIR29B2CHG*_C3** ***PDE4C*_C3** ***ASPA*_C1** ***FHL2*_C10** ***KLF14*_C1**	**98.72%**	**0.988**	**0.996**
Test tissue (n = 192)	**M8**	**8**	***FHL2*_C4** ***TRIM59*_C7** ***KLF14*_C3**	**97.4%**	**0.968**	**0.991**

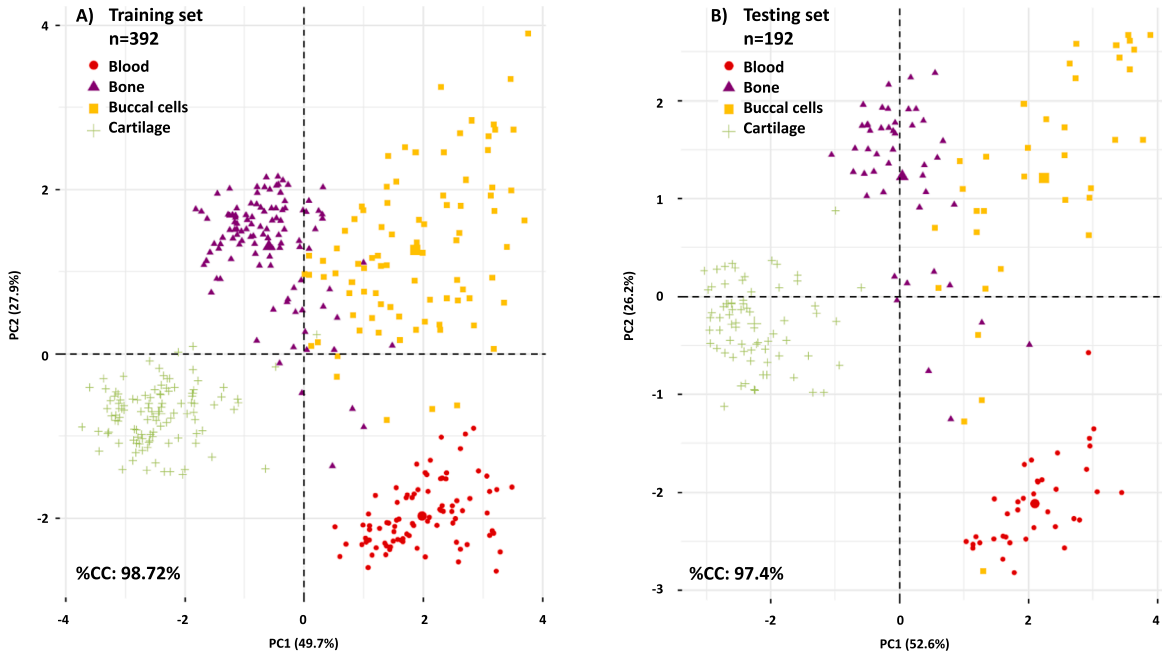


Fig. 2. Principal Component Analysis (PCA) based on eight CpGs (*EDARADD*_C1, *TRIM59*_C1, *ELOVL2*_C6, *MIR29B2CHG*_C3, *PDE4C*_C3, *ASPA*_C1, *FHL2*_C10 and *KLF14*_C1) for A) the training (n = 392: 109 cartilage samples, 110 bone samples, 87 blood samples and 86 buccal swabs) and B) the testing set (n = 192: 72 cartilage samples, 46 bone samples, 38 blood samples and 36 buccal swabs). Blood samples, buccal swabs, bone, and cartilage are depicted as red, orange, violet and green datapoints, respectively.

4. Discussion

Age estimation using epigenetic markers has been widely explored during the last ten years [40]. The observation of high correlation of DNA methylation levels with chronological age has been translated into a large list of age prediction models. Since DNA methylation is tissue-dependent, models are usually developed based on specific tissues. Indeed, to date, age estimation has been mainly focused on the most commonly analyzed forensic tissues such as blood [41,42], buccal cells [43], saliva [44], semen [45,46] or nails [47]. Nevertheless, skeletal remains are also important sources of biological material for both forensic and anthropological analyses. To address such DNA analyses, several studies have introduced age prediction models for teeth and bone [48–50]. Although at the moment multiple age prediction models can be found in the literature, few provide coverage for various tissues at the same time. Ten years ago, Horvath reported the first pan-tissue epigenetic clock applicable to 51 tissues and cell types [28], five of them forensically relevant: blood, buccal cells, saliva, epidermis and knee cartilage. However, the Horvath model needs DNA methylation data for 353 CpG sites that is not useful when working with DNA samples of poor quality and quantity common to forensic analysis. In order to solve this constrain, in 2019, Jung et al. introduced the first forensic tissue-combined model for age inference covering three somatic tissues – blood, saliva and buccal swabs – using capillary electrophoresis [31]. Two years later, the VISAGE Enhanced Tool was developed based on massively parallel sequencing, designed to estimate age from blood and buccal cells, as well as bone samples [32].

Regarding additional forensic specimens affected by decomposition, an initial attempt to infer age was recently undertaken by Becker et al. [51]. In this study, age estimation for elastic cartilage from the epiglottis was explored using a protein clock based on racemization of aspartic acid and accumulation of pentosidine, which provided an MAE of ± 4 years. Moreover, costal cartilage samples were also explored using the age prediction model developed for bone generated with the VISAGE Enhanced Tool [32]. Nevertheless, high prediction errors were obtained (MAE: ± 25.8 years for cartilage samples compared to ± 3.4 years for bone) [32], thus justifying our development of an age prediction model specifically for cartilage.

To the best of our knowledge, this is the first study reporting forensic age estimation for cartilage based on DNA methylation. To establish an efficient predictive model, the VISAGE Enhanced Tool comprising a total of 44 CpG sites located in eight genes (*KLF14*, *TRIM59*, *MIR29B2CHG, FHL2*, *ELOVL2*, *EDARADD, PDE4C* and *ASPA*) has been explored in 181 costal cartilage samples representing a full adult age range [32]. From these candidates, three CpG sites from three genes were selected to build the cartilage-specific age prediction model using multivariate quantile regression (*FHL2*_C4, *TRIM59*_C7 and *KLF14*_C3). Cross-validation of the model provided an MAE of ± 4.41 years and RMSE of 5.52. When assessing this tool in an independent test set, comparable results were obtained (MAE: ± 4.26 years, RMSE: 5.39), which were also similar to the values obtained from the protein clock-based system reported by Becker et al. (MAE: ± 4.0 years) [51]. However, prediction errors for the cartilage model were slightly higher if comparing with other epigenetic clocks generated from the VISAGE Enhanced Tool for blood (MAE: ± 3.2 years), buccal cells (MAE: ± 3.7 years) and bone (MAE: ± 3.4 years) [32]. When plotting the predicted age against chronological age (Fig. 1), a slight overestimation of age in young samples as well as an underestimation in older ones was observed, especially in the test set, which indicates the utility of age-specific prediction intervals compared to the predicted age. Although both training and test sets have similar age range distributions, this effect could be explained by a reduced sample size (n = 109 and n = 72, for training and test sets, respectively). However, due to the nature of the specimens, broadly-based sample collection for a complete age range is difficult to achieve.

As a supplementary analysis to age estimation, identification of the biological source of a stain using DNA methylation analysis becomes a useful application in forensic genetics. The initial development of the VISAGE Enhanced Tool for age prediction from somatic tissues allowed the inference of the epigenetic age of somatic tissues such as blood, buccal cells and bone [32], which is now extended to cartilage samples. To broaden the applications of this MPS tool, a tissue prediction model was additionally developed. DNA methylation has previously shown its potential for tissue identification based on the fact that this epigenetic signature affects gene expression and therefore, genomic loci are

differentially methylated between tissues. However, selection of previous markers was exclusively made to accomplish tissue differentiation [52–54]. Moreover, assessment of such assays was based on the detection of the marker presenting the methylated signal, leading to a direct assignment to the corresponding tissue of origin. In the present model, a quantitative rather than a qualitative evaluation was applied. Therefore, the novelty of this assay consists not only in simultaneous use of the VISAGE Enhanced Tool for inference of age and tissue source, but on developing a different approach for the latter. From the 44 CpG sites analyzed, eight markers were selected to build the tissue prediction model using multinomial logistic regression (*EDARADD*_C1, *TRIM59*_C1, *ELOVL2*_C6, *MIR29B2CHG*_C3, *PDE4C*_C3, *ASPA*_C1, *FHL2*_C10 and *KLF14*_C1), classifying blood, buccal cells, bone, and cartilage. Although not all the eight markers contributed to a full separation of the four tissues, cross-validation of the 8-CpG combined model provided 98.72% correct classifications, with high sensitivity (0.988) and specificity (0.996) values. The assessment of an independent testing set led to similar results of 97.4% of correct classifications, 0.968 for sensitivity and 0.991 for specificity.

To date, methods for body fluid identification have largely relied on chemical or immunological tests [55]. However, lack of specificity and requirements for large amounts of sample related to these methods can be solved using epigenetic markers. In this way, through a single reaction, differentiation of four potential tissue sources can be achieved. Since it is relevant to detect the presence of body fluids or tissues in biological specimens, differentiation between blood and buccal cells can add a determinant value to the forensic investigations. However, differentiation between cartilage and bone is much less critical for forensic purposes, since sample collection determines per se the origin of the sample. Differences between cartilage and bone in *EDARADD* have been previously found in other animal species, such as Baboons [56]. Although from a forensic point of view, the differentiation of cartilage and bone is largely unnecessary; we note that epigenetic differences between both tissues could be useful in clinical applications when studying the development of degenerative diseases such as osteoarthritis [57].

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgements

This study received support from the European Union's Horizon 2020 Research and Innovation Program under grant agreement No. 740580 within the framework of the Visible Attributes through Genomics (VISAGE) Project and Consortium and form an institutional grant for young scientists from the Medical University of Silesia in Katowice, Poland (grant no. PCN-2-053/K/2/O).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102936.

References

[1] M.D. Vigeland, T. Egeland, Joint DNA-based disaster victim identification, Sci. Rep. 11 (2021), 13661.
[2] M. Prinz, A. Carracedo, W.R. Mayr, N. Morling, T.J. Parsons, A. Sajantila, et al., DNA commission of the international society for forensic genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI), Forensic Sci. Int. Genet. 1 (1) (2007) 3–12.
[3] G. Calacal, F. Delfin, M. Tan, L. Roewer, D. Magtanong, M. Lara, et al., Identification of exhumed remains of fire tragedy victims using conventional methods and autosomal/Y-chromosomal short tandem repeat DNA profiling, Am. J. Forensic Med. Pathol. 26 (3) (2005) 285–291.
[4] A. Ossowski, M. Kuś, P. Brzeziński, J. Prüffer, J. Piątek, G. Zielińska, et al., Example of human individual identification from World War II gravesite, Forensic Sci. Int. 233 (1–3) (2013) 179–192.
[5] M. Baeta, C. Núñez, S. Cardoso, P.-M. L, L. Herrasti, F. Etxeberria, et al., Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship, Forensic Sci. Int. Genet. 19 (2015) 272–279.
[6] M. Alvarez-Cubero, M. Saiz, L. Martinez-Gonzalez, J. Alvarez, A. Eisenberg, B. Budowle, et al., Genetic identification of missing persons: DNA analysis of human remains and compromised samples, Pathobiology 79 (5) (2012) 228–238.
[7] A. Nadeem, M. Ashraf, N. Qadeer, K. Rizwan, A. Mehmood, A. AlZahrani, et al., Tracking missing person in large crowd gathering using intelligent video surveillance, Sensors 22 (14) (2022) 5270.
[8] National Crime Information Center - FBI, Criminal Justice Information Service (CJIS), Missing Pers. Unidentified Pers. Stat. (2021). ⟨https://www.fbi.gov/file-repository/2021-ncic-missing-person-and-unidentified-person-statistics.pdf/view⟩.
[9] https://www.interpol.int/News-and-Events/News/2021/INTERPOL-unveils-new-global-database-to-identify-missing-persons-through-family-DNA.
[10] M. Tomsia, K. Droździok, G.T. Javan, R. Skowronek, M. Szczepański, E. Chełmecka, Costal cartilage ensures low degradation of DNA needed for genetic identification of human remains retrieved at different decomposition stages and different postmortem intervals, Post. Hig. i Med Doświadczalnej / Adv. Hyg. Exp. Med 75 (2021) 852–858.
[11] M. Tomsia, Kornelia Droździok, P. Banaszek, M. Szczepański, A. Pałasz, E. Chełmecka, The intervertebral discs' fibrocartilage as a DNA source for genetic identification in severely charred cadavers, Forensic Sci. Med Pathol. 18 (2022) 442–449.
[12] J. Becker, N. Mahlke, S. Ritz-Timme, P. Boehme, The human intervertebral disc as a source of DNA for molecular identification, Forensic Sci. Med. Pathol. 17 (4) (2021) 660–664. Available from: papers2://publication/uuid/AA4D3970-F698-408D-9644-5827F2A2141B.
[13] M.L. Goff, Early postmortem changes and stages of decomposition, Exp. Appl. Acarol. 49 (1–2) (2009) 21–36.
[14] T. Siriboonpiputtana, T. Rinthachai, J. Shotivaranon, V. Peonim, B. Rerkamnuaychoke, Forensic genetic analysis of bone remain samples, Forensic Sci. Int. 284 (2018) 167–175.
[15] J. Jakubowska, A. Maciejewska, R. Pawłowski, Comparison of three methods of DNA extraction from human bones with different degrees of degradation, Int. J. Leg. Med. 126 (1) (2012) 173–178.
[16] Y. Seo, D. Uchiyama, K. Kuroki, T. Kishida, STR and mitochondrial DNA SNP typing of a bone marrow transplant recipient after death in a fire, Leg. Med. 14 (6) (2012) 331–335.
[17] Sanz-Piña, E. Santurtún, A. Zarrabeitia, M. Sanz-Piña, et al., Forensic Sci. Int 2019 (2019) 302, 109862.
[18] M. Tomsia, J. Ciesla, Joanna Pilch-Kowalczyk, Przemysław Banaszek, E. Chełmecka, Cartilage tissue in forensic science—state of the art and future research directions, Processes 10 (2022) 2456.
[19] M. Tomsia, M. Głaz, J. Nowicka, M. Szczepanski, Sodium nitrite detection in costal cartilage and vitreous humor – Case report of fatal poisoning with sodium nitrite, J. Forensic Leg. Med. 81 (2021), 102186.
[20] M. Tomsia, J. Nowicka, R. Skowronek, M. Woś, J. Wójcik, K. Droździok, et al., A comparative study of ethanol concentration in costal cartilage in relation to blood and urine, Processes 8 (2020) 1637.
[21] M. Tomsia, J. Nowicka, R. Skowronek, G.T. Javan, E. Chełmecka, Concentrations of volatile substances in costal cartilage in relation to blood and urine – preliminary studies, Arch. Med. Sadowej Kryminol. 71 (1–2) (2021) 38–46.
[22] J. Hardy, S. Chrosciany, J. Bernard, C. Mabit, P. Marcheix, The human costal cartilage: Anatomical and radiological study of macro-vascularization and micro-vascularization and its clinical relevance regarding vascularized chondrocostal free flap surgery, Ann. Anat. 232 (2020), 151581.
[23] E. Safronova, N. Borisova, S. Mezentseva, K. Krasnopol'skaya, Characteristics of the macromolecular components of the extracellular matrix in human hyaline cartilage at different stages of ontogenesis, Biomed. Sci. 2 (2) (1991) 162–168.
[24] H. Pfeiffer, H. Mörnstad, A. Teivens, Estimation of chronologic age using the aspartic acid racemization method. I. On human rib cartilage, Int. J. Leg. Med. 108 (1) (1995) 19–23.
[25] T. Ikeda, Estimating age at death based on costal cartilage calcification, Tohoku J. Exp. Med. 243 (4) (2017) 237–246.
[26] S. Zhang, J. Zhen, H. Li, S. Sun, H. Wu, P. Shen, et al., Characteristics of Chinese costal cartilage and costa calcification using dual-energy computed tomography imaging, Sci. Rep. 7 (2017), 2923.
[27] J. Fernández-Tajes, A. Soto-Hermida, M. Vázquez-Mosquera, E. Cortés-Pereira, A. Mosquera, M. Fernández-Moreno, N. Oreiro, C. Fernández-López, et al., Genome-wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients, Ann. Rheum. Dis. 73 (4) (2014) 668–677.
[28] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (10) (2013) R115.
[29] K. Alkass, B.A. Buchholz, S. Ohtani, T. Yamamoto, H. Druid, K.L. Spalding, Age estimation in forensic sciences: application of combined aspartic acid racemization and radiocarbon analysis, Mol. Cell Proteom. 9 (5) (2010) 1022–1030.
[30] A. Pisarek, E. Pośpiech, A. Heidegger, C. Xavier, A. Papież, D. Piniewska-Róg, et al., Epigenetic age prediction in semen - marker selection and model development, Aging 13 (15) (2021) 19145–19164.
[31] 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.
[32] A. Woźniak, A. Heidegger, D. Piniewska-róg, E. Pośpiech, A. Pisarek, E. Kartasińska et al., Development of the VISAGE Enhanced Tool and statistical models for epigenetic age estimation in blood, buccal cells and bones 13 5 2021.
[33] B. Koop, F. Mayer, T. Gündüz, J. Blum, J. Becker, J. Schaffrath, et al., Postmortem age estimation via DNA methylation analysis in buccal swabs from corpses in different stages of decomposition-a "proof of principle" study, Int J. Leg. Med 135 (1) (2021) 167–173.
[34] D.R. Masser, A.S. Berg, W.M. Freeman, Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing, Epigenetics Chromatin 6 (1) (2013), 33.
[35] R. Koenker, S. Portnoy, P.T. Ng, A. Zeileis, P. Grosjean, C. Moler et al., Quantile Regres., Package "quantreg. " 2019.
[36] H. Wickham, W. Chang, Creat. Elegant Data Vis. Using Gramm. Graph., Package "ggplot2. " 2019.
[37] B. Ripley, W. Venables, Feed-Forw. Neural Netw. Multinomial Log. -Linear Models, Package "nnet. " 2022.
[38] A. Kassambara, F. Mundt, Extr. Vis. Results Multivar. Data Anal., Package "factoextra. " 2022.
[39] A. Alfons Cross-Valid. tools Regres. Models, Package "cvTools. " 2015.
[40] R. Noroozi, S. Ghafouri-Fard, A. Pisarek, J. Rudnicka, M. Spólnicka, W. Branicki, et al., DNA methylation-based age clocks: from age prediction to age reversion, Ageing Res Rev. 68 (2021), 101314.
[41] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Z. Makowska, A. Pałeczka, et al., Development of a forensically useful age prediction method based on DNA methylation analysis, Forensic Sci. Int. Genet. 17 (2015) 173–179.
[42] A. Freire-Aradas, L. Girón-Santamaría, A. Mosquera-Miguel, A. Ambroa-Conde, C. Phillips, M. Casares de Cal, et al., A common epigenetic clock from childhood to old age, Forensic Sci. Int. Genet. 60 (2022), 102743.
[43] A. Ambroa-Conde, L. Girón-Santamaría, A. Mosquera-Miguel, C. Phillips, M. A. Casares de Cal, A. Gómez-Tato, et al., Epigenetic age estimation in saliva and in buccal cells, Forensic Sci. Int. Genet. 61 (2022), 102770.
[44] 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.
[45] 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.
[46] A. Heidegger, A. Pisarek, M. de la Puente, H. Niederstätter, E. Pośpiech, A. Woźniak, et al., Development and inter-laboratory validation of the VISAGE enhanced tool for age estimation from semen using quantitative DNA methylation analysis, Forensic Sci. Int. Genet. 56 (2022), 102596.
[47] K. Fokias, L. Dierckx, W. Van de Voorde, B. Bekaert, Age determination through DNA methylation patterns in fingernails and toenails, Forensic Sci. Int. Genet. 16 (64) (2023), 102846.
[48] 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 (10) (2015) 922–930.
[49] H. Correia Dias, L. Manco, F. Corte Real, E. Cunha, A blood–bone–tooth model for age prediction in forensic contexts, Biology 10 (12) (2021) 1312.
[50] H.Y. Lee, S.R. Hong, J.E. Lee, I.K. Hwang, N.Y. Kim, J.M. Lee, et al., Epigenetic age signatures in bones, Forensic Sci. Int. Genet. 46 (2020).
[51] J. Becker, N.S. Mahlke, A. Reckert, S.B. Eickhoff, S. Ritz-Timme, Age estimation based on different molecular clocks in several tissues and a multivariate approach: an explorative study, Int. J. Leg. Med. 134 (2) (2020) 721–733.
[52] H. Lee, J. An, S. Jung, Y. Oh, E. Lee, A. Choi, et al., Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers, Forensic Sci. Int. Genet. 17 (2015) 17–24.
[53] D. Silva, J. Antunes, K. Balamurugan, G. Duncan, C. Alho, B. McCord, Developmental validation studies of epigenetic DNA methylation markers for the detection of blood, semen and saliva samples, Forensic Sci. Int. Genet. 23 (2016) 55–63.
[54] H. Holtkötter, V. Beyer, K. Schwender, A. Glaub, K. Johann, M. Schürenkamp, et al., Independent validation of body fluid-specific CpG markers and construction of a robust multiplex assay, Forensic Sci. Int. Genet. 29 (2017) 261–268.
[55] T. Sijen, S. Harbison, On the identification of body fluids and tissues: a crucial link in the investigation and solution of crime, Genes 12 (11) (2021) 1728.
[56] G. Housman, E.E. Quillen, A.C. Stone, An evolutionary perspective of DNA methylation patterns in skeletal tissues using a baboon model of osteoarthritis, J. Orthop. Res. 39 (10) (2021) 2260–2269.
[57] Jb van Meurs, C. Boer, L. Lopez-Delgado, J. Riancho, Role of epigenomics in bone and cartilage disease, J. Bone Min. Res. 34 (2) (2019) 215–230.

Artículo 4: Inference of tobacco and alcohol consumption habits from DNA methylation analysis of blood

A. Ambroa-Conde, M. A. Casares de Cal, A. Gómez-Tato, O. Robinson, A. Mosquera-Miguel, M. de la Puente, J. Ruiz-Ramírez, C. Phillips, M. V. Lareu, A. Freire-Aradas.

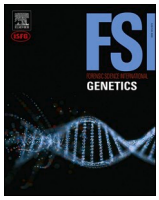
Forensic Science International: Genetics, volumen 70, página 103022, 2024 DOI: 10.1016/j.fsigen.2024.103022

Forensic Science International: Genetics 70 (2024) 103022

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



Inference of tobacco and alcohol consumption habits from DNA methylation analysis of blood

A. Ambroa-Conde [a], M.A. Casares de Cal [b], A. Gómez-Tato [b], O. Robinson [c], A. Mosquera-Miguel [a], M. de la Puente [a], J. Ruiz-Ramírez [a], C. Phillips [a], M.V. Lareu [a], A. Freire-Aradas [a,*]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, Universidade de Santiago de Compostela, Spain
[b] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain
[c] MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

ARTICLE INFO

Keywords:
DNA methylation
Logistic regression
Quantile regression
Blood
Tobacco
Alcohol
Age estimation

ABSTRACT

DNA methylation has become a biomarker of great interest in the forensic and clinical fields. In criminal investigations, the study of this epigenetic marker has allowed the development of DNA intelligence tools providing information that can be useful for investigators, such as age prediction. Following a similar trend, when the origin of a sample in a criminal scenario is unknown, the inference of an individual's lifestyle such as tobacco use and alcohol consumption could provide relevant information to help in the identification of DNA donors at the crime scene. At the same time, in the clinical domain, prediction of these trends of consumption could allow the identification of people at risk or better identification of the causes of different pathologies. In the present study, DNA methylation data from the UK AIRWAVE study was used to build two binomial logistic models for the inference of smoking and drinking status. A total of 348 individuals (116 non-smokers, 116 former smokers and 116 smokers) plus a total of 237 individuals (79 non-drinkers, 79 moderate drinkers and 79 drinkers) were used for development of tobacco and alcohol consumption prediction models, respectively. The tobacco prediction model was composed of two CpGs (cg05575921 in *AHRR* and cg01940273) and the alcohol prediction model three CpGs (cg06690548 in *SLC7A11*, cg0886875 and cg21294714 in *MIR4435–2HG*), providing correct classifications of 86.49% and 74.26%, respectively. Validation of the models was performed using leave-one-out cross-validation. Additionally, two independent testing sets were also assessed for tobacco and alcohol consumption. Considering that the consumption of these substances could underlie accelerated epigenetic ageing patterns, the effect of these lifestyles on the prediction of age was evaluated. To do that, a quantile regression model based on previous studies was generated, and the potential effect of tobacco and alcohol consumption with the epigenetic age was assessed. The Wilcoxon test was used to evaluate the residuals generated by the model and no significant differences were observed between the categories analyzed.

1. Introduction

DNA methylation has become a biomarker of interest in the forensic field. It has been studied for individual age estimation [1,2], tissue determination [3] and differentiation of monozygotic twins [4]. Additionally, since this marker undergoes changes caused by exogenous agents [5] and medical disorders [6], its use has been proposed for the study of environmental factors and diseases in the clinical domain [7]. DNA methylation has also been correlated with lifestyle factors in both clinical [8–10] and forensic [11,12] fields. In relation to the clinical context, development of indices based on epigenetic markers as risk indicators in health disorders related to tobacco and alcohol could be of great interest. In the case of tobacco, an index based on buccal cells has been developed allowing discrimination between normal and cancerous cells [13,14]. For alcohol, a correlation between methylation and consumption disorders has so far been demonstrated [15]. Prediction of these trends of consumption could allow the identification of people at risk or a better identification of the causes of different pathologies. In forensic DNA analysis, the prediction of lifestyles and environmental exposures would allow a better characterization of unknown

* Corresponding author.
E-mail address: ana.freire@usc.es (A. Freire-Aradas).

https://doi.org/10.1016/j.fsigen.2024.103022
Received 22 August 2023; Received in revised form 22 December 2023; Accepted 25 January 2024
Available online 28 January 2024

perpetrators from a biological trace. Therefore, the inference of tobacco and alcohol consumption from DNA could be used as a DNA-intelligence tool for police investigations helping to reduce the number of suspects when a DNA sample cannot be matched to any reference sample or profile stored in DNA databases [12].

In recent years, several articles have been published assessing DNA methylation differences between tobacco and alcohol users and non-users. In the case of tobacco, a tendency towards hypomethylation has been observed in smokers compared to non-smokers [16,17]. However, in the case of alcohol, although clear differences have been observed in the methylation patterns of drinkers and non-drinkers [15], no clear methylation trend has been observed. Different studies have identified general tendencies for drinkers to hypermethylate [18,19], and non-drinkers to hypomethylate [20–22] or show both [23]. It has been shown that these differentially methylated patterns between tobacco and alcohol users and non-users can be reversed at some positions [20, 24,25]. For alcohol consumption, a partial recovery of methylation levels was observed after several weeks of abstinence [20]. For tobacco, on the other hand, a more in-depth study identified reversible and irreversible positions over time after cessation of smoking [25–27]. For some positions, the methylation values of former smokers, recovered to levels observed in non-smokers within 0 to 35 years [25–27]. In contrast, other CpG positions maintained methylation levels of smokers even 35 years after cessation [25].

As a result of the discovery articles published to date, many DNA methylation markers (CpG positions) correlated with the consumption of both substances have been identified, and prediction models of smoking and alcohol status built accordingly. In the case of tobacco, the models generated have been developed mainly with blood samples, using different population groups, technologies and statistical models [28–32]. In the model generated by Elliott et al. [28] the differences between two populations were directly assessed (Europe vs South Asia). It was observed that for some CpG positions, the methylation values differed greatly between the groups analyzed, but it was possible to generate a Random Forest model with sensitivity and specificity values of more than 80% for both populations for predicting current active smokers. Furthermore, additional logistic regression models were used to infer years since smoking cessation, number of cigarette per day, as well as years as a smoker [30,33]. Finally, using a single marker, Philibert et al. [31] generated prediction models for blood and saliva while assessing sex and age as covariables.

For the prediction of drinking status, a smaller number of models have been developed to date [32,34–36], all of them blood-based and mainly for European populations. It is worth mentioning the model developed by Liu et al. composed of 144 CpGs obtained Area Under the Curve (AUC) values for heavy drinkers vs non-drinkers of 0.80 to 0.99 in four replication cohorts [34]. For this model, subsequent studies evaluated a possible overestimation of the results by overfitting, obtaining AUC values between 0.50 and 0.75 [36–38]. Further evaluation of this model is necessary to avoid the potential overestimation observed. In addition, two other alcohol prediction models were developed obtaining AUC values of 0.73 [35] and 0.74 [32] for the classification of light to moderate vs heavy drinker, and non-drinker vs heavy drinker, respectively.

The study of such lifestyle factors in the forensic field is of interest not only for the inference of consumption by itself, but also to evaluate their effect on the individual age prediction models developed so far. It has been shown that the consumption of alcohol and tobacco may be correlated with age acceleration, as they are contributing factors in age-related diseases [39]. Age acceleration caused by these lifestyle factors was initially assessed and some correlation was observed, but further research is needed [39–41].

In the present study, DNA methylation data from the UK AIRWAVE study [42] were used to generate binomial logistic regression models for the classification of smoking and drinking alcohol status. The database used was generated from the Airwave Health Monitoring Study [43], that has been recruiting participants among UK police officers since 2004. These models were generated using 348 individuals and 237 individuals for tobacco and alcohol consumption, respectively. In the case of tobacco, a final model comprising 2-CpGs was selected that addressed non-smokers + former smokers vs active smokers. For alcohol, however, non-drinker and moderate drinker were grouped together to be compared with heavy drinkers, selecting a final model comprising 3-CpGs. Finally, age prediction models were assessed showing they were not affected by an individual's smoking or drinking status.

2. Material and methods

2.1. DNA methylation data and sample classification

DNA methylation data was accessed from the UK AIRWAVE study [42]. A total of 1115 blood samples (452 females and 663 males) were evaluated, which had been analyzed with the Infinium MethylationEPIC BeadChip array, composed of 853,307 CpGs. The age range of the samples analyzed was 19 to 65 years old (standard deviation: ± 13.52 years). For these samples, data related to their lifestyle were available in the form of a questionnaire. Questions relating to tobacco and alcohol consumption were used to classify the samples into different categories.

For smoking status classification, the questions evaluated were: whether or not the volunteer is a smoker, and if currently they do not smoke but have smoked five or more cigarettes a day in the past. Taking into account the first question, individuals who answered "NO" were classified as non-smokers and those who answered "YES" were classified as smokers. The individuals that had answered "YES" to the second question were classified as former smokers. After performing this classification, the individuals were grouped as following: 116 smokers, 728 non-smokers and 271 former smokers (detailed information can be found in Table 1). For model building, the number of individuals in each group was matched, taking as reference the category with the lowest number of classified individuals (116 smokers).

Table 1
Summary of the tobacco and alcohol group classification for the 1115 blood samples.

Lifestyle	Group	Sample size	Gender	Age (years old)
Tobacco	Non-smoker	728	281 women 447 men	19-65
	Smoker	116	45 women 71 men	20-65
	Former smoker	271	126 women 145 men	22-64
Alcohol	Non-drinker	79	44 women 35 men	21-65
	Moderate drinker	956	364 women 592 men	19-65
	Heavy drinker	79	44 women 35 men	20-64

For alcohol classification, five questions referring to the consumption of different alcoholic beverages for one week were considered. For this purpose, the total amount of alcoholic units consumed by each participant was assessed, classifying them into three groups based on this value. For a correct classification, the gender of the participants was considered. Therefore, the classification was performed according to the following conditions: values equal to 0 as non-drinker; values > 0 and $\leq$ 14 units per week for women as moderate drinkers; values > 0 and ≤ 21 units for men as moderate drinkers; values > 14 units for women as heavy drinkers; and values > 21 units for men as heavy drinkers. From this classification, the individuals were grouped as following: 79 non-drinkers, 956 moderate drinkers and 80 heavy drinkers (detailed information is shown in Table 1). For model building, the number of individuals of the groups was matched, with reference to the category with the lowest number of classified individuals (79 non-drinkers and heavy drinkers).

2.2. Statistical analysis

The selection of candidate CpG sites to infer the studied lifestyles was based on the AUC (Area under the ROC Curve), and on the percentage of correct classifications (%CC), both obtained from binomial logistic regression analysis. Firstly, AUC was calculated for all the 853,307 CpGs included in the Infinium MethylationEPIC BeadChip array using the pROC R package [44], and those presenting values equal or higher to 0.7 were retained. For these retained CpGs and taking into account the maximum number of CpGs recommended to be included in logistic regression models without overfitting – $p + 1 \leq min\ (n0, n1, n2)/10$ *parameters* [45] – %CC was calculated and those markers depicting values equal or higher to 70% were selected. Statistical significance was set at p-value ≤ 0.05 (E-2).

For the lifestyle prediction models two different statistical approaches were used, binomial and multinomial logistic models, developed using the nnet package [46] for multinomial regression. The corresponding predictive accuracy for the logistic regression models was measured with the following performance metrics: sensitivity, specificity, AUC and percentage of correct classifications (%CC). For the evaluation of these parameters in binomial logistic models, it should be considered that smoker and heavy drinker groups were set as class 1, so a better prediction of belonging to this group produces a higher sensitivity. For the other group assessed in each model, class 0, the specificity will be the parameter related to its classification. Principal Component Analysis (PCA) was carried out using the factoextra R package [47]. Cross-validation of the developed logistic regression models was performed with a leave-one-out cross validation using the pROC R package.

An age prediction model based on multivariate quantile regression was built using the quantreg R package [48] taking as reference a previous model [49]. Cross-validation of the age prediction model was performed with a k-fold cross-validation ($k = 10$) using cvTools R package [50]. For the age prediction model, the median absolute error (MAE) was used to measure the predictive accuracy. Representations of the DNA methylation values, as well as the predicted vs chronological age were made using the ggplot2 R package [51]. Correlations between DNA methylation levels and chronological age were evaluated using the Spearman correlation test (r_s), and the inter-group variability was analyzed using the standard deviation (SD). All statistical analyses were carried out using R software v.4.1.1 [52] with scripts developed in-house.

3. Results

3.1. Selection of candidate CpGs

For the selection of the markers correlated with tobacco and alcohol, matched samples for age and sex for a total of 116 non-smokers (age range: 19–62, mean: 40.56 years; female/male ratio: 0.55) vs 116 smokers (age range: 20–65, mean: 41.73 years; female/male ratio: 0.63), plus 79 non-drinkers (age range: 21–65, mean: 42.13 years; female/male ratio: 1.26) vs 79 drinkers (age range: 20–64, mean: 42.65 years; female/male ratio: 1.19) were assessed. A total of 853,307 binomial logistic regression models were generated (one per CpG included in the Infinium MethylationEPIC BeadChip array). Of all the evaluated models, those with an AUC value below 0.7 were discarded, keeping a total of 67 tobacco-correlated CpGs and 30 alcohol-correlated CpGs, that were selected for further evaluation.

In order to reduce the number of markers analyzed, the formula $p + 1 \leq min\ (n0, n1, n2)/10$ *parameters* [53] was used in order to define the maximum number of CpGs recommended to be included in logistic regression models without overfitting. Thus, the number of events per variable was evaluated based on the number of individuals, taking into account that a minimum of 10 events per parameter is advisable to avoid overfitting [45,53]. The number of markers that could be used depends on the number of parameters in the model, with binomial models (dichotomous variables) allowing a larger number of markers than multinomial models (polytomous variables). Therefore, with the evaluated groups defined as *n0*, *n1* and *n2* using the previous formula [45], for tobacco, with a maximum of 116 individuals in one of the evaluated groups (smokers), the generated models should not contain more than 4 and 10 CpGs for multinomial and binomial logistic models, respectively. For alcohol, both non-drinker and heavy drinker groups contained an equal number of 79 individuals, reducing the number of recommended markers to 3 and 6 for multinomial and binomial models, respectively. Since many CpGs presented very similar AUC values and considering that a forward approach was used according to the CpGs sorted first by AUC and then by %CC, it was decided to select as many markers as allowed in the binomial models for each lifestyle. For this final selection, following the order of markers defined by the AUC, percentage of correct classifications was calculated and only those CpGs, up to the maximum CpG number, as defined above, with a value equal to or higher than 70% of correct classifications were selected.

Based on these criteria and avoiding those CpGs that had more than 50 'not analyzed' (NA), a total of 10 tobacco-correlated and 5 alcohol-correlated CpGs were selected. Figs. 1 and 2 show the corresponding boxplots for the DNA methylation values per group for tobacco and alcohol, respectively. Although selection of candidate CpGs was performed using the extreme groups for both cases, boxplots have been arranged to show the DNA methylation levels for the three groups per lifestyle. Table 2 shows the selected markers for both lifestyles ordered by the criteria defined above. As it can be observed, all the selected markers (10 and 5 CpGs for tobacco and alcohol, respectively) were statistically significant (p-values range between E-15 to E-5), although showing the tobacco-correlated CpGs higher correlation values than the corresponding CpGs for alcohol (mean: E-12 versus E-6, respectively).

Potential correlation with age for the 15 selected CpG sites was evaluated for each category inside each lifestyle, with all markers showing r_s values below or close to $|0.50|$ (Supplementary Table S1). When assessing the correlation with age, generally higher r_s values were observed for the smoker group (mean: $|0.27|$) compared to non-smokers (mean: $|0.12|$) and former smokers (mean: $|0.14|$). In the case of alcohol, the moderate and heavy drinker groups showed similar mean r_s values (mean of $|0.26|$ and $|0.29|$, respectively), with non-drinkers giving lower values for the majority of markers (mean: $|0.08|$). It is noteworthy that differences were observed in some specific markers, e.g., *SLC7A11* and *MIR4435–2HG*, with mainly one category standing out from the rest (-0.46 for heavy drinkers and -0.34 for moderate drinkers, respectively). In addition, evaluation of the dispersion of DNA methylation values of the groups, indicated similar trends observed between the groups of the studied lifestyles (tobacco presenting mean values for non-smokers of 0.04, former smokers: 0.05 and smokers: 0.06; alcohol presenting for all groups a mean value of 0.04).

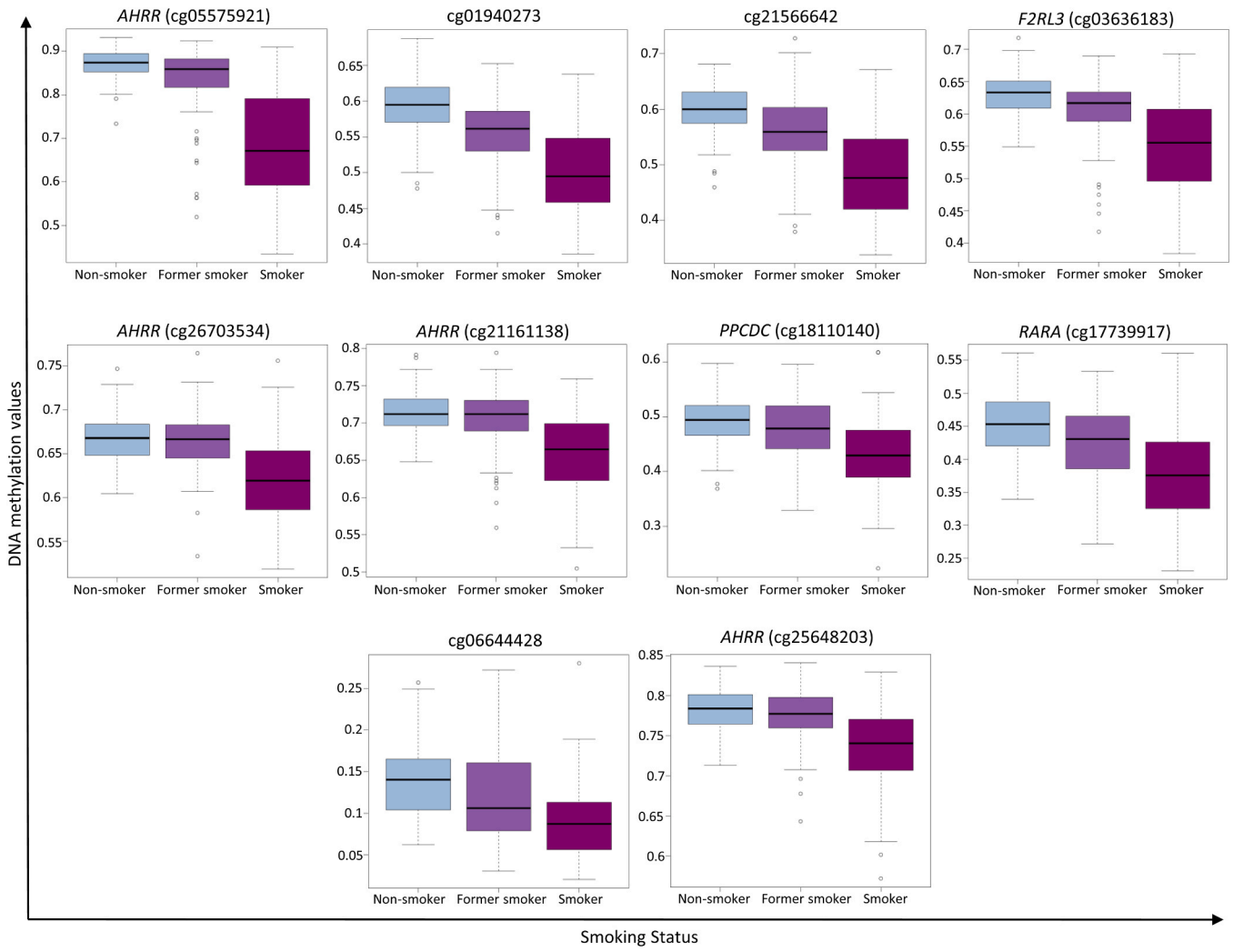


Fig. 1. Boxplots representing the DNA methylation values for the 10 CpGs correlated with tobacco consumption presenting an AUC and %CC higher than 0.7 and 70%, respectively, for N = 116 non-smoker, N = 116 former smoker and N = 116 smoker individuals. Markers are sorted in descending order by AUC and %CC.

3.2. *Development of a prediction model for tobacco consumption*

A multinomial logistic regression model was explored to differentiate three categories for smoking status (N = 116 smokers, N = 116 former smokers and N = 116 non-smokers). For this evaluation, several combinations of the selected CpG sites were tested. Considering the maximum number of markers recommended for the tobacco multinomial models (4 CpGs), four combinations of markers were evaluated following a forward approach, i.e. the addition of one marker per subsequent model. The addition of CpG sites stopped when no additional improvement was discernible from the model. The corresponding percentage of correct classifications is shown in Table 3. Following the order established for the selected CpGs (Table 2), models were generated including one by one, the CpGs presenting the highest AUC values in descending order. As shown in Table 3, a slight increase in the global correct classification rate was observed, as the number of markers in the models increased (from 58.91% to 66.09%, for the 1-CpG through to the 4-CpGs model, respectively). In detailed assessments of the specific categories, we observed that the extreme groups gave higher classification rates (mean: 73.28% and 70.48% for non-smokers and smokers, respectively) compared with former smokers (mean: 46.55%).

Considering the results achieved, a multinomial logistic model does not readily provide correctly classified consumption habits for the three groups under study. Therefore, binomial logistic regression was subsequently explored. To perform these analyses while keeping all the samples from the three categories represented, it was decided to group two categories into one. For this purpose, the classification table (Table 3) of the most accurate multinomial model (4-CpGs) was used to assess the classification trend of the intermediate group (former smokers). In the multinomial model, 52.59% of former smokers were correctly classified, with most of the remaining individuals (36.21%) being classified as non-smokers. Consequently, between these two groups, 88.80% of the former smokers were present, with a greater tendency to be classified as non-smokers than smokers (36.21% and 11.21%, respectively). Therefore, for the creation of the binomial logistic models, it was decided to combine non-smoker and former smoker as a single category. Additionally, in order to avoid losing informativeness, all the samples from both groups were retained. Therefore, binomial logistic models were generated by comparing the non-smokers + former smokers group of 232 individuals with 116 smokers. Following the same approach as the one used for multinomial models, up to three models were assessed for this analysis. The corresponding performance metrics are shown in Table 4. In this case, since no improvement in the classifications was achieved by the third CpG included in the model, the models tested stopped at three CpG sites, and a 2-CpG model of *AHRR* (cg05575921) and cg01940273 was selected (Fig. 3A), providing 86.49% of correct classifications. The additional performance metrics included an AUC level of 0.87, while specificity (0.90) was higher than sensitivity (0.79), indicating a better classification of the non-smokers + former smokers group vs smokers (%CC: 90.09% vs 79.31%, respectively).

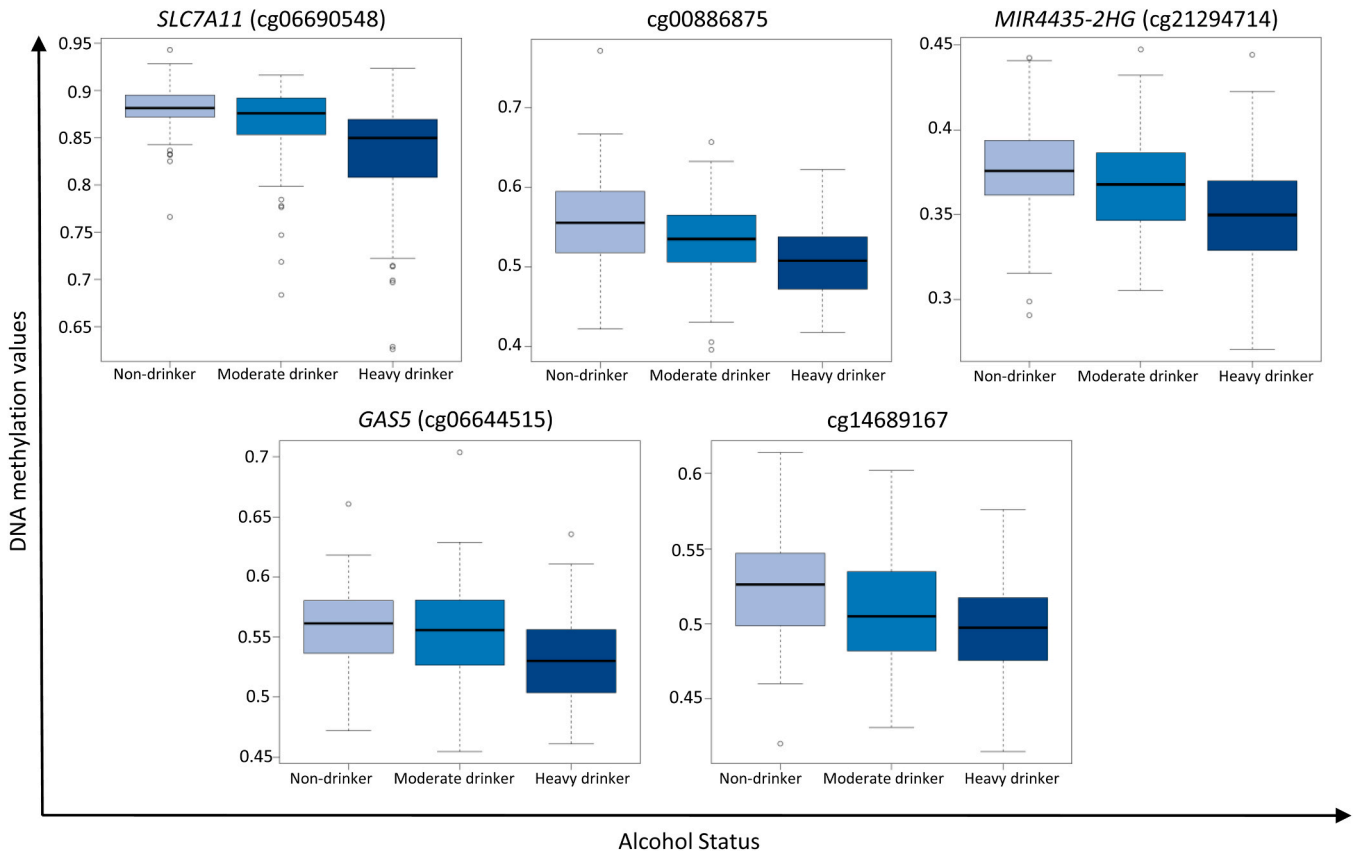


Fig. 2. Boxplots representing the DNA methylation values for the 5 CpGs correlated with alcohol consumption presenting an AUC and %CC higher than 0.7 and 70%, respectively, for N = 79 non-drinker, N = 79 moderate drinker and N = 79 drinker individuals. Markers are sorted in descending order by AUC and %CC.

Table 2
Preliminary selection of 10 smoking and 5 drinking related CpGs showing a value higher than 0.7 for AUC, percentage of correct classifications (%CC) and statistical significance (p-value), based on DNA methylation data from the DPUK platform.

Lifestyle	Gene	CpG_ID	GRCh38 chromosome position	AUC	%CC	P-value
Tobacco	*AHRR*	cg05575921	chr5:373263	0.90	88.79%	1.77E-11
	none	cg01940273	chr2:232420224	0.89	86.21%	3.73E-15
	none	cg21566642	chr2:232419951	0.89	83.19%	9.57E-15
	F2RL3	cg03636183	chr19:16889774	0.86	81.47%	1.71E-12
	AHRR	cg26703534	chr5:377243	0.82	76.29%	1.27E-12
	AHRR	cg21161138	chr5:399245	0.81	77.59%	3.99E-12
	PPCDC	cg18110140	chr15:75058039	0.81	73.71%	1.30E-11
	RARA	cg17739917	chr17:40321320	0.80	76.72%	3.12E-12
	none	cg06644428	chr2:232419402	0.80	72.84%	1.92E-11
	AHRR	cg25648203	chr5:395329	0.79	73.28%	5.67E-11
Alcohol	*SLC7A11*	cg06690548	chr4:138241654	0.82	77.85%	3.44E-7
	none	cg00886875	chr3:106635478	0.76	70.89%	4.84E-7
	MIR4435-2HG	cg21294714	chr2:111429551	0.74	72.15%	4.63E-6
	GAS5	cg06644515	chr1:173865693	0.73	70.89%	5.64E-6
	none	cg14689167	chr10:4267023	0.71	71.52%	1.92E-5

Considering the reduced number of samples, a leave-one-out cross-validation was performed to validate the selected model. The cross-validation showed the same values for sensitivity, specificity and %CC as the model (0.79, 0.90 and 86.49%, respectively), with only a slight difference in the AUC value obtained (0.86).

As no further data for smokers was available, the discarded individuals from the non-smokers and former smokers group were used as a testing set to assess the accuracy of the model for this category. Thus, tobacco consumption status was predicted for a total of 606 non-smokers and 151 former smokers, obtaining a 93.39% of correctly classified non-smokers and former smokers vs only 6.61% classified as smokers.

3.3. Development of a prediction model for alcohol consumption

To explore multinomial logistic regression for alcohol consumption, a similar strategy was used to that outlined above for tobacco. DNA methylation values available for 79 non-drinkers, 79 moderate drinkers and 79 heavy drinkers were used to generate a drinking status model. Models were generated following a forward approach, adding in descending order the CpGs listed in Table 2. The maximum number of variables (3 CpGs) considering that groups had 79 individuals was taken into account and subsequently, a maximum of three models were tested. The corresponding performance metrics are detailed in Table 3. The multinomial models generated for the prediction of drinking status

Table 3
Summary of the accuracy of the evaluated multinomial logistic regression models for tobacco (N = 116 non-smoker, N = 116 former smoker and N = 116 smoker), and alcohol status prediction (N = 79 non-drinker, N = 79 moderate drinker and N = 79 heavy drinker).

	Model	CpGs	Correct Classifications	Classification Table			
Tobacco	1-CpG model	*AHRR* (cg05575921)	58.91%	**Predicted\Real**	**Non-smoker**	**Former smoker**	**Smoker**
				Non-smoker	74.14%	52.59%	12.93%
				Former smoker	25.00%	32.76%	17.24%
				Smoker	0.86%	14.66%	69.83%
	2-CpGs model	*AHRR* (cg05575921), cg01940273	64.37%	**Predicted\Real**	**Non-smoker**	**Former smoker**	**Smoker**
				Non-smoker	72.41%	37.07%	11.21%
				Former smoker	25.86%	50.86%	18.97%
				Smoker	17.24%	12.07%	69.83%
	3-CpGs model	*AHRR* (cg05575921), cg01940273, cg21566642	64.37%	**Predicted\Real**	**Non-smoker**	**Former smoker**	**Smoker**
				Non-smoker	73.28%	37.07%	11.21%
				Former smoker	25.86%	50.00%	18.97%
				Smoker	0.86%	12.93%	69.83%
	4-CpGs model	*AHRR* (cg05575921), cg01940273, cg21566642, *F2RL3*	66.09%	**Predicted\Real**	**Non-smoker**	**Former smoker**	**Smoker**
				Non-smoker	73.28%	36.21%	10.34%
				Former smoker	25.86%	52.59%	17.24%
				Smoker	0.86%	11.21%	72.41%
Alcohol	1-CpG model	*SLC7A11*	48.52%	**Predicted\Real**	**Non-drinker**	**Moderate drinker**	**Heavy drinker**
				Non-drinker	70.89%	56.96%	21.52%
				Moderate drinker	21.52%	20.25%	24.05%
				Heavy drinker	7.59%	22.78%	54.43%
	2-CpGs model	*SLC7A11*, cg00886875	50.21%	**Predicted\Real**	**Non-drinker**	**Moderate drinker**	**Heavy drinker**
				Non-drinker	65.82%	43.04%	17.73%
				Moderate drinker	22.78%	26.58%	24.05%
				Heavy drinker	11.39%	30.38%	58.23%
	3-CpGs model	*SLC7A11*, cg00886875, *MIR4435-2HG*	52.74%	**Predicted\Real**	**Non-drinker**	**Moderate drinker**	**Heavy drinker**
				Non-drinker	63.29%	49.37%	12.66%
				Moderate drinker	25.32%	27.85%	21.52%
				Heavy drinker	11.39%	22.78%	65.82%

showed low percentages of correct classifications (mean: 50.49%). The highest %CC was obtained for the 3-CpGs model (52.74%), with a large proportion of the misclassified individuals been related to the moderate drinker category (> 70%).

The results obtained for multinomial logistic regression do not allow an adequate classification of the three categories under study. Therefore, binomial logistic regression was then evaluated. As in the case of tobacco, it was decided to combine two of the categories, while retaining all the samples from the three groups represented. When evaluating the classification table of the most accurate multinomial model (3-CpGs), the classification trend of the intermediate group was used (moderate drinkers). It was observed that 49.37% of the moderate drinkers were classified as non-drinkers, with the combination of these categories accounting for the majority of the observed moderate drinkers (77.22%). Therefore, it was decided, considering the trend of the intermediate group, to combine non-drinkers and moderate drinkers into a single category. Additionally, to avoid exclusion of individuals used in marker selection, we grouped non-drinkers and moderate drinkers adding all individuals of each group. Thus, binomial logistic regression models were built for 158 individuals of the non-drinkers + moderate drinkers group against 79 heavy drinkers. The forward approach was used for model building and the corresponding performance metrics evaluated (Table 4). In this case, a 3-CpG model composed of *SLC7A11* (cg06690548), cg0886875 and *MIR4435–2HG* (cg21294714) was selected (Fig. 3B), providing 74.26% correct classifications. Additional performance metrics comprised an AUC level of 0.80 and sensitivity (0.81) was higher than specificity (0.71), reflecting a better classification of the heavy drinker group vs non-drinkers + moderate drinkers (%CC: 81.01% vs 70.89%, respectively).

To evaluate the accuracy of the selected model, a leave-one-out cross-validation was carried out showing similar values to those obtained in the selected model (AUC: 0.79, sensitivity: 0.80, specificity: 0.70 and %CC: 72.57%).

In the case of alcohol, the remaining individuals in the available dataset were moderate drinkers (N = 858), with any non-drinkers and/or heavy drinkers available for testing. It was decided to use these individuals as a testing set to assess the accuracy of the model for such a heterogeneous group as moderate drinkers. Of the 858 individuals available, 852 (99.30%) were correctly classified as non-drinkers + moderate drinkers, and only 6 individuals (0.7%) were classified as heavy drinkers.

3.4. Tobacco and alcohol effects on age prediction

To evaluate the effects of the lifestyle factors assessed in this study on the prediction of epigenetic age, a quantile regression model was developed. This model was performed using the DNA methylation data from the same sample set used for building the tobacco and alcohol prediction models, except for three samples which presented missing data for the CpGs under study in this section (N = 1112, 19 to 65 years old, analyzed with Infinium MethylationEPIC BeadChip array). Based on the markers employed in a previous publication [49], five out of the seven reported CpGs were used for this analysis: *ELOVL2* (cg21572722), *ASPA* (cg02228185), *FHL2* (cg06639320), *CCDC102B* (cg19283806)

Table 4
Summary of the accuracy of the evaluated binomial logistic regression models for tobacco (N = 232 non-smoker + former smoker vs N = 116 smoker), and alcohol status prediction (N = 158 non-drinker + moderate drinker vs N = 79 heavy drinker). The selected final models are highlighted in bold.

	Model	CpGs	AUC	Sensitivity	Specificity	Correct Classifications	Classification Table		
Tobacco	1-CpG model	AHRR (cg05575921)	0.87	0.81	0.87	85.06%	**Predicted\Real**	**Non-smoker + Former smoker**	**Smoker**
							Non-smoker + Former smoker	87.07%	18.97%
							Smoker	12.93%	81.03%
	2-CpGs model	**AHRR (cg05575921), cg01940273**	**0.87**	**0.79**	**0.90**	**86.49%**	**Predicted\Real**	**Non-smoker + Former smoker**	**Smoker**
							Non-smoker + Former smoker	90.09%	20.69%
							Smoker	9.91%	79.31%
	3-CpGs model	AHRR (cg05575921), cg01940273, cg21566642	0.87	0.79	0.90	86.21%	**Predicted\Real**	**Non-smoker + Former smoker**	**Smoker**
							Non-smoker + Former smoker	89.66%	20.69%
							Smoker	10.34%	79.31%
Alcohol	1-CpG model	SLC7A11	0.77	0.70	0.77	74.26%	**Predicted\Real**	**Non-drinker + Moderate drinker**	**Heavy drinker**
							Non-drinker + Moderate drinker	76.58%	30.38%
							Heavy drinker	23.42%	69.62%
	2-CpGs model	SLC7A11, cg00886875	0.78	0.75	0.70	71.31%	**Predicted\Real**	**Non-drinker + Moderate drinker**	**Heavy drinker**
							Non-drinker + Moderate drinker	69.62%	25.32%
							Heavy drinker	30.38%	74.68%
	3-CpGs model	**SLC7A11, cg00886875, MIR4435-2HG**	**0.80**	**0.81**	**0.71**	**74.26%**	**Predicted\Real**	**Non-drinker + Moderate drinker**	**Heavy drinker**
							Non-drinker + Moderate drinker	70.89%	18.99%
							Heavy drinker	29.11%	81.01%
	4-CpGs model	SLC7A11, cg00886875, MIR4435-2HG, *GAS5*	0.80	0.81	0.70	73.84%	**Predicted\Real**	**Non-drinker + Moderate drinker**	**Heavy drinker**
							Non-drinker + Moderate drinker	70.25%	18.99%
							Heavy drinker	29.75%	81.01%

and cg07082267. The scatter plots of the corresponding age-correlated CpGs representing the DNA methylation values against the chronological age are shown in Supplementary Fig. S1. Quantile regression was performed using the five selected CpGs in order to generate the age prediction model, giving a median absolute error (MAE) of ± 2.8 years. To evaluate the accuracy of the model, a k-fold cross-validation was carried out, providing a MAE of ± 2.79 years.

In order to check if the studied lifestyles could have an influence on age prediction, the samples used for building the age prediction model were grouped by categories in the same way as in the final logistic regression models, and errors per cluster were calculated. For the samples categorized by tobacco intake, the predicted errors obtained for the two groups of interest were evaluated, resulting in a MAE of ± 2.78 years for the non-smokers + former smokers category and a MAE of ± 3.12 years for the smokers category (Fig. 4A). To assess whether there were significant differences between the residuals of the evaluated groups (Fig. 4B), a Wilcoxon test was applied and a p-value of 0.78 obtained. Hence, no significant differences in age predictions were observed for non-smokers or former smokers and smokers.

Regarding the alcohol intake, the predicted errors obtained for the groups of interest were also evaluated, yielding a MAE of ± 2.82 years for the non-drinker + moderate drinker category and a MAE of ± 2.59 years for the heavy drinker group (Fig. 4C). To determine if there were significant differences between the residuals of the groups studied (Fig. 4D), a Wilcoxon test was carried out and a p-value of 0.10 obtained. Therefore, no significant differences were observed between the predicted ages of non-drinkers or moderate drinkers and heavy drinkers.

4. Discussion

DNA methylation has become an increasingly studied biomarker of interest in forensic genetics. In recent years, the different applications of this epigenetic marker have been evaluated, highlighting the prediction of age [1,2], tissue identification [3], and the study of lifestyles [11,12]. Tobacco and alcohol consumption have created the most interest, and several discovery articles have been published identifying markers that show differences between consumers and non-consumers [21,22,26,27]. As these behaviors are related to disease risk/status, the methylation levels in certain genes correlated with medical conditions such as cancer have been evaluated, and many show differences between smokers or drinkers, and non-consumers [8,9]. At the same time, associations have been observed between dependency and DNA methylation, therefore, the generation of predisposition indices could be useful for preventative diagnostics [13,14]. From these studies, tools of forensic interest have been generated to identify whether a person is a smoker or non-smoker and a drinker or non-drinker [31,32,36] - information that could be

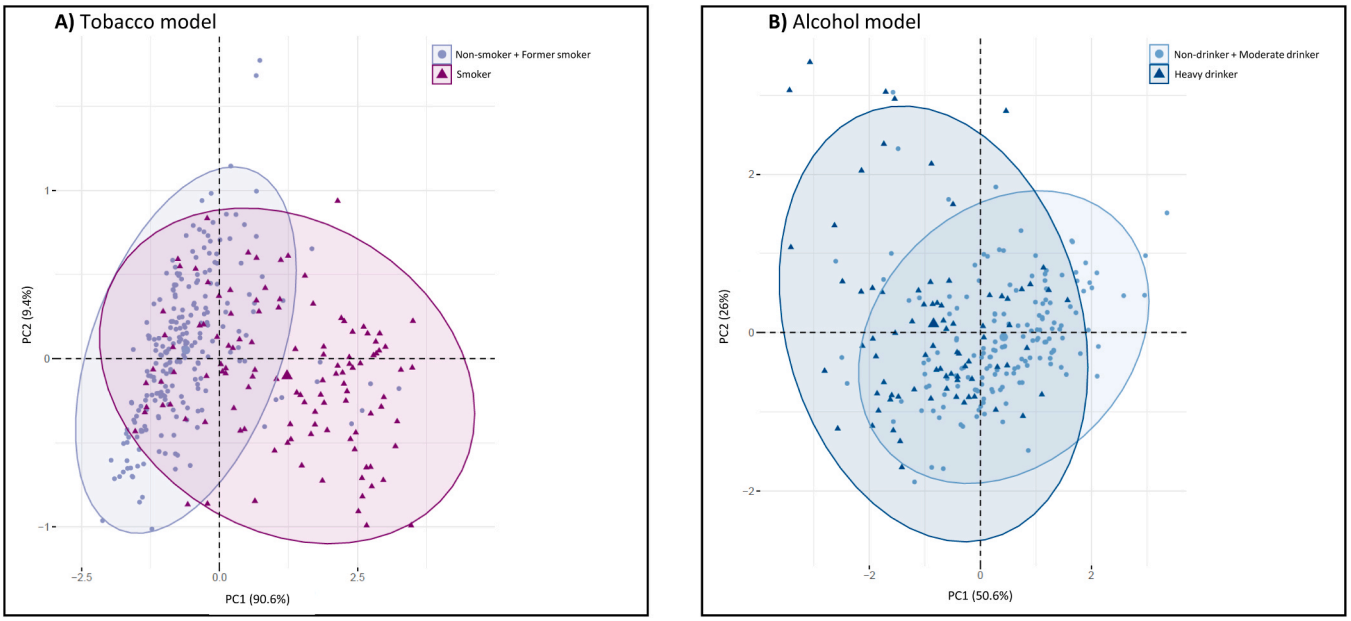


Fig. 3. PCA representation of the selected models for each lifestyle. **A)** Tobacco binomial logistic regression model composed of *AHRR* and cg01940273 representing 232 individuals classified as non-smoker + former smoker as well as 116 smokers. **B)** Alcohol binomial logistic regression model composed of *SLC7A11,* cg0886875 and *MIR4435–2HG* representing 158 individuals classified as non-drinker + moderate drinker as well as 79 heavy drinkers.

relevant in criminal investigations. Finally, the relationship between these lifestyle habits and age acceleration has been evaluated, causing epigenetic modifications of biological age to differ more from the chronological age than under normal conditions [39–41]. Considering both clinical and forensic applications, markers correlated with smoking and drinking status were selected to develop predictive models for these lifestyles.

In the present study, a total of 15 markers were selected, 10 CpGs for smoking and 5 CpGs for alcohol consumption. Regarding the underlying DNA methylation data, it is important to mention that inter-individual variation within each smoking or alcohol category was observed, with the most marked variation found in consumers, i.e. former and current smokers, as well as moderate and heavy drinkers. Similar findings have been previously reported by Vidaki et al. [54], suggesting a potential consumer-behaviour effect in terms of intensity and duration.

To identify the most accurate consumption status prediction models among the 10 tobacco-related CpGs and the 5 alcohol-related CpGs selected in our study, a forward approach was explored using logistic regression. Considering the three categories defined for each lifestyle (non-consumer, intermediate consumer, consumer), multinomial and binomial models were evaluated to represent the different consumption states analyzed. As Maas et al. [36] mentioned in reference to overfitting, for all the models generated in our study, the recommendations for the amount of predictors per number of participants were taken into account. To avoid overfitting, the generation of the models was limited to a maximum number of CpGs depending on whether the analysis was multinomial or binomial [45,53]. Firstly, the multinomial logistic regression models were evaluated, presenting percentages of correct classifications of 66.09% for the 4-CpG tobacco model and 52.74% for the 3-CpG alcohol model. As observed in other published models, the intermediate groups show lower accuracies than the extreme ones. In the case of tobacco consumption, Alghanim et al. [29] developed a 4-CpG multinomial logistic regression model that gave high percentages of correct classifications for non-smokers (84.9%) and smokers (90%), although these percentages were reduced to 66.7% for former smokers. At the same time, the model of 13-CpGs published by Maas et al. [30] shows sensitivity and specificity values (sensitivity: 0.78 for non-smokers, 0.65 for former smokers and 0.67 for smokers; specificity: 0.75 for non-smokers, 0.77 for former smokers and 0.99 for smokers) similar to the 4-CpG multinomial model of our study (sensitivity: 0.73 for non-smokers, 0.53 for former smokers and 0.72 for smokers; specificity: 0.77 for non-smokers, 0.78 for former smokers and 0.94 for smokers). The sensitivity for former smokers, which evaluates the degree of correct prediction of this category, presents values lower than 0.7 in both the Maas et al. model (0.65) and from our study (0.53). For the alcohol model developed at the present study, the percentages of correct classifications were lower (63.29% for non-drinkers, 27.85% for moderate drinkers and 65.82% for heavy drinkers), with the intermediate group being the most challenging to be classified (0.28 sensitivity).

Considering the difficulty observed in predicting the three groups separately, grouping of the intermediate category with one of the extreme groups was evaluated. For tobacco, grouping of non-smokers and former smokers was previously considered by the models of Elliott et al. [28] and Maas et al. [30], obtaining high values for the statistical parameters evaluated in those models. Moreover, in the 1-CpG binomial logistic regression models developed by Alghanim et al. [29], a reduction in AUC was observed for non-smokers vs former smokers compared to smokers vs former smokers in blood (mean AUC 0.73 and 0.91, respectively) and in saliva (mean AUC 0.69 and 0.81, respectively). Thus, former smokers were differentiated better from smokers than from non-smokers. In the case of drinking status, Maas et al. [36] evaluated different clustering of categories in the models analyzed, with the highest AUCs obtained for the models comparing heavy drinkers vs non-drinkers and light drinkers (AUC > 0.7 for the different CpG combinations evaluated). Evaluation of the grouping of consumers (heavy + moderate) vs non-consumers was evaluated by Chamberlain et al. [32], obtaining in their model an AUC of 0.64. With all this, it could be concluded that, if a multinomial model for these lifestyles does not correctly predict the three categories, it would be advisable to evaluate the grouping of the intermediate category with the non-consumers.

Although the grouping of two categories (non-consumers and intermediate consumption) could be considered a disadvantage, in a forensic context, the value of a test could lie in identifying the group that is less frequent in the general population. Bearing this in mind, according to Eurostat data, only 19.7% of the EU population smokes daily and only 8.4% drinks daily. In the case of alcohol, almost one in five individuals are considered heavy drinkers, showing differences related either to the gender (26.3% for men and 11.4% for women), and to the country of

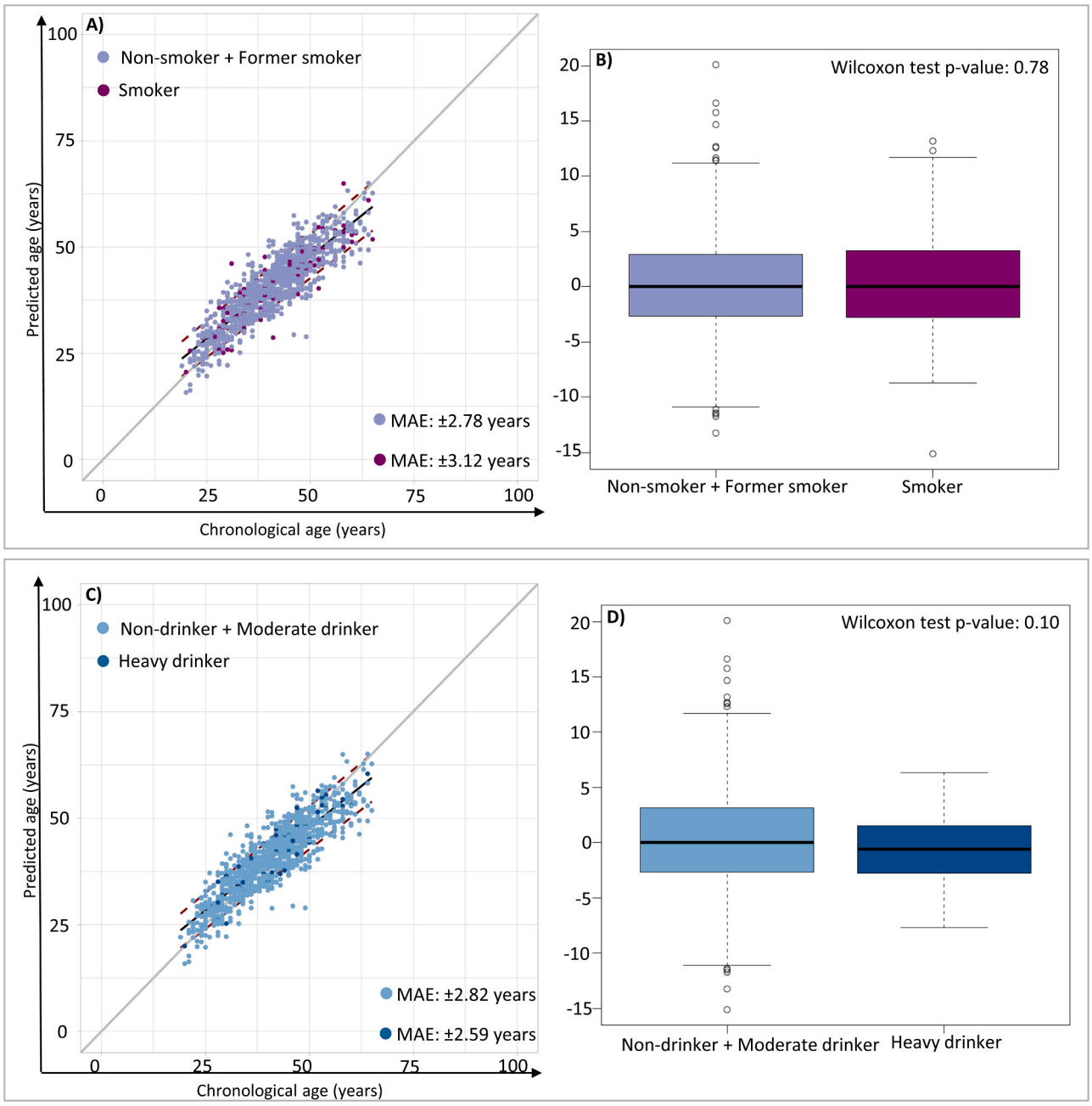


Fig. 4. Graphical representations derived from the age prediction quantile regression model generated with 1112 blood samples previously used for building lifestyle prediction models. **A)** and **C)** Predicted age vs chronological age of the sample set grouped by tobacco and alcohol categories, respectively. The black diagonal line represents the 0.5 quantile and the discontinuous dark red lines the corresponding 0.1 and 0.9 quantiles. The gray line represents perfect correlation. **B)** and **D)** Residuals obtained in the model generated for the defined categories for tobacco and alcohol consumption, respectively.

origin, showing Denmark and Romania the highest percentage of prevalence of heavy episodic drinking at least once a month among alcohol drinkers (37.8% and 35.0%, respectively). Therefore, correctly identifying individuals who might fall into these categories would further reduce the number of suspects in a criminal case.

For tobacco consumption, a 2-CpG binomial logistic regression model confronting non-smokers + former smokers vs smokers composed of *AHRR* (cg05575921) and cg01940273 gave performance metrics of: AUC: 0.87, sensitivity: 0.79, specificity: 0.90 and %CC: 86.49%. To validate the model using an independent testing set, since a larger pool of samples was not available, a complete external validation was not performed, evaluating only individuals classified as non-smoker (N = 606) and former smoker (N = 151). This allowed evaluation of the efficiency of the model for the pooled group, with 93.39% of the validation set samples correctly assigned. Although these results demonstrate that a correct classification of the non-smokers + former smokers was obtained, further analysis of the models would be necessary to fully evaluate a larger validation set covering all three evaluated categories.

Most published models make a partial evaluation of the categories, evaluating two of three possible groups. Only those models built with the extreme categories (non-smokers vs smokers) have given AUC values close to or above 0.90 [32,55–57]. However, biased results could be generated, since former smokers are not represented.

A very comprehensive model is described by Maas et al. [30], reporting a binomial logistic model (grouping non-smoker and former smoker in one category) giving an AUC of 0.90 and a multinomial model

with AUCs of 0.84, 0.77 and 0.93 for non-smoker, former smokers and smokers, respectively. Our study has certain similarities with Maas et al., but the different maker selection criteria and differences in model construction (backward and forward approaches) led to the generation of very different prediction models. The model developed in our study, built with 2 CpGs, that are among the 13 selected by Maas et al., gave similar results with a lower number of markers (AUC: 0.87). It is worth mentioning that there is a large difference in the sensitivity of both models, with that of Maas et al. predicting smokers with a value of 0.59 compared to our model with 0.79. A possible justification for the high AUC and correct classifications obtained by Maas et al. despite the low sensitivity could be a large imbalance in the number of individuals among the groups in the model. The Maas et al. multinomial model presented difficulties to predict former smokers, obtaining sensitivities of 0.65 for the model and 0.39 for the validation set. The prediction of the intermediate category has been more challenging, with worse predictions being observed in the models that evaluate it independently (Shenker et al. [58] AUC: 0.82, sensitivity: 0.69 and specificity: 0.71; Maas et al. [30] AUC: 0.77, sensitivity: 0.65 and specificity: 0.77). Considering the reversibility of DNA methylation for some of the markers correlated with tobacco and the effect on this reversibility of time since cessation of smoking and intensity of consumption demonstrated by McCartney et al. [33], the group of former smokers is a difficult category to classify consistently. Moreover, as shown in Fig. 3A, and in the study of McCartney et al., it is difficult to obtain a complete separation of the categories analyzed, as there are several factors that can modify the methylation patterns associated with this lifestyle. More accurate prediction of the intermediate group is likely to require development of: a larger number of markers; specific markers for former smokers (e.g., positions that exhibit irreversible methylation patterns over time); or complementing the generated models with consumption cessation time models.

The selected tobacco markers for the logistic regression model for predicting smoking status have been previously reported. The *AHRR* gene is a protein-coding gene related to cell growth and differentiation. The CpG position selected in this study, cg05575921, has been reported on multiple occasions as one of the markers that shows the highest correlation with tobacco consumption [25–27,59] and is present in almost all smoking status prediction models published to date [28, 30–32,55–57,60]. Some of these obtained individual AUCs for this marker of 0.88 [30] and 0.99 [31], similar results to our 1-CpG model generated with this marker (non-smokers vs smokers of 0.90, and non-smokers + former smokers vs smokers of 0.87). The *AHRR* marker is of interest for not presenting differences between European and South Asian populations [28], for recovering methylation values to a non-smoking state after 5 years of consumption cessation [17], for being correlated with age acceleration [9] and for being correlated with different mortality factors [14] - representing as it does a biomarker of lung cancer [10]. Position cg01940273 has also been previously reported to correlate with smoking status [25–27]. This position is present in the 13 CpG model of Maas et al. [30] presenting individual AUCs of 0.89, similar to those of our 1-CpG model for non-smokers vs smokers (0.89). Maas's study also evaluated the time since smoking cessation, a characteristic with which this marker had been related in other publications [14]. Finally, it has been observed that cg01940273 is related to breast cancer risk [22,23].

Evaluating the markers correlated with alcohol consumption status, a 3-CpG prediction model for non-drinkers + moderate drinkers vs heavy drinkers composed of *SLC7A11* (cg06690548), cg0886875 and *MIR4435–2HG* (cg21294714) gave an AUC of 0.80, sensitivity of 0.81, specificity of 0.71 and %CC of 74.26%. Different approaches to clustering the categories in binomial models have been evaluated in the published studies for alcohol. Those models composed only of non-drinkers vs heavy drinkers generally have higher values, presenting AUCs around 0.80 [32,34,36]. Of the published models, only Liu et al. [34] and Maas et al. [36] attempt to assess all possible categories for alcohol consumption. Focusing on Maas's model, AUCs in a range of 0.70–0.75 were obtained for heavy drinkers vs non-drinkers + light drinkers, lower than those obtained with our model with similar grouping (0.80). Considering the classification methods, the model for heavy + at risk drinkers vs non-drinkers + light drinkers could also be compared with our study. The classification of alcohol consumption presents a greater challenge than tobacco consumption, with consistently lower AUC and %CC values, as well as a smaller separation between the defined categories, as can be seen in Fig. 3B. This might be due to a smaller separation in the methylation values of the individuals classified in the different categories evaluated, with the intermediate group overlapping with the two extreme categories to a greater extent (Fig. 2). This could also be observed in the classification tables of the multinomial models generated, with the predictions of the intermediate group more evenly distributed than former smokers in the tobacco models. Therefore, more studies are needed to obtain a better separation of the intermediate group for which the introduction of other variables such as time or intensity of consumption could be useful.

Of the markers correlated with alcohol consumption used in the selected model, *SLC7A11*, has been reported in lifestyle-related discovery studies. For the other selected CpGs, to the best of our knowledge, our study detected these CpG positions to be correlated with alcohol consumption for the first time. The *SLC7A11* gene, is the most reported marker related to alcohol consumption, its correlation has been observed in different discovery studies [21,22] and it is among the selected CpGs for all the previous published drinking status models. Moreover, this marker has been associated with the number of drinks consumed per week, and may be a marker of interest to identify heavy or at-risk drinkers [22]. Both cg21294714 of the *MIR4435–2HG* gene and cg0886875, were identified to be correlated with alcohol consumption in this study. These CpG positions presented individually, for the extreme categories (non-drinkers vs heavy drinkers), gave AUC values of 0.70 and 0.71 respectively.

One limitation in our study is the absence of a complete independent sample set for both smoking and alcohol consumption. Additionally, both CpG selection and model building were performed using identical samples and this could partially bias our findings. However, the resulting selected and most informative CpG sites are consistent with previous studies, which adds value to our results. For those CpG sites reported in the present study as correlated with alcohol for the first time will, additional studies will be necessary for validation purposes.

Different publications have shown that tobacco and alcohol consumption influence age acceleration, causing discordance between chronological and biological age data. When developing age prediction models based on DNA methylation, it is often not possible to check whether the selected markers are influenced by other factors. Indeed, the present study has been developed using samples from the UK AIRWAVE study, where the volunteers were police officers. Since this occupational group is exposed to high levels of stress, a potential cofounding effect cannot be discarded in our analyses. Considering that these lifestyle factors produce global alterations in DNA methylation values, an age prediction model based on quantile regression was generated to evaluate the effect of the consumption of these substances in age prediction. A MAE of ± 2.79 years was obtained. It should be noted that in the study of Freire-Aradas et al. [49], the MAE obtained was ± 3.07. The difference in the evaluated parameter is probably due to the range of age analyzed. While Freire-Aradas's model covers 18 to 104 years, the dataset used in the current study covers 19 to 65 years old. The Wilcoxon test indicated no significant differences (tobacco p-value of 0.78; alcohol p-value of 0.10) between the residuals generated by the model for the groups analyzed in each lifestyle. Although an effect of smoking and alcohol consumption on age prediction was not observed, this only means that the markers included in the age prediction model developed on the present study are not sensitive to these lifestyle factors. This cannot be taken to mean that smoking and alcohol do not have an effect on aging in general.

Additionally, in accordance with Vidaki et al., 2023 [54], signs of association with age were found for the majority of smoking-CpGs depicting methylation reduction through the lifespan, although all r_s values were below or close to |0.50|. From these, the highest values were observed for the smoker group in our study. Similar findings were obtained for the alcohol-CpGs, although in this case, no specific group was detected to present higher correlations in comparison to the others. Although the correlation with age is low in scale, it has been detected and therefore, the age predictive potential of the selected CpGs should be further studied.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgements

MVL is supported by the Ministerio de Educación, Cultura y Ciencia, Spain (PID2019-107876RB-I00). MdlP is supported by a postdoctoral fellowship awarded by the Gobierno de España: IJC2020-042638-I, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU/PRTR". J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039). The Airwave Health Monitoring Study is funded by the Medical Research Council (MRC), (MR/R023484/1), the National Institute for Health Care Research (NIHR) Health Protection Research Unit in Chemical and Radiation Threats and Hazards (NIHR-200922), the Imperial College Biomedical Research Centre (BRC) 2017–22, and the Imperial College Healthcare NHS Trust. The initial phase of the study, including participant recruitment, was funded by the Home Office (780-TETRA; 2003-18). Views expressed are those of the authors and not necessarily those of the study sponsors. We thank all study participants for their involvement. DPUK provided data access for this project: Elliott, P. (2017). Airwave [Data set]. Dementias Platform UK. https://doi.org/10.48532/002000 through MRC grant ref MR/L023784/2″ (core funding).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2024.103022.

References

[1] K. Schwender, O. Holländer, S. Klopfleisch, M. Eveslage, M.F. Danzer, H. Pfeiffer, et al., Development of two age estimation models for buccal swab samples based on 3 CpG sites analyzed with pyrosequencing and minisequencing, Forensic Sci. Int Genet 53 (2021) 102521.
[2] A. Wózniak, A. Heidegger, D. Piniewska-Róg, E. Pospiech, C. Xavier, A. Pisarek, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones, Aging (Albany NY) 13 (5) (2021) 6459–6484.
[3] H.Y. Lee, S.E. Jung, E.H. Lee, W.I. Yang, K.J. Shin, DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood, Forensic Sci. Int Genet 24 (2016) 75–82.
[4] A. Vidaki, C. Díez López, E. Carnero-Montoro, A. Ralf, K. Ward, T. Spector, et al., Epigenetic discrimination of identical twins from blood under the forensic scenario, Forensic Sci. Int Genet 31 (2017) 67–80.
[5] A.V. Probst, E. Dunleavy, G. Almouzni, Epigenetic inheritance during the cell cycle, Nat. Rev. Mol. Cell Biol. 10 (3) (2009) 192–206.
[6] C.B. Santos-Rebouças, M.M.G. Pimentel, Implication of abnormal epigenetic patterns for human diseases, Eur. J. Hum. Genet 15 (1) (2007) 10–17.
[7] K.M. Bakulski, M.D. Fallin, Epigenetic epidemiology: promises for public health research, Environ. Mol. Mutagen 55 (3) (2014) 171–183.
[8] M. Varela-Rey, A. Woodhoo, M.L. Martinez-Chantar, J.M. Mato, S.C. Lu, Alcohol, DNA methylation, and cancer, Alcohol Res Curr. Rev. 35 (1) (2012) 25–35.
[9] X. Gao, Y. Zhang, L.P. Breitling, H. Brenner, Tobacco smoking and methylation of genes related to lung cancer development, Oncotarget 7 (37) (2016) 59017–59028.
[10] D. Fragou, E. Pakkidi, M. Aschner, V. Samanidou, L. Kovatsi, Smoking and DNA methylation: Correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure, Food Chem. Toxicol. 129 (2019) 312–327.
[11] H.Y. Lee, S.D. Lee, K.J. Shin, Forensic DNA methylation profiling from evidence material for investigative leads, BMB Rep. 49 (7) (2016) 359–369.
[12] A. Vidaki, M. Kayser, From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence, Genome Biol. 18 (1) (2017) 1–13.
[13] A.E. Teschendorff, Z. Yang, A. Wong, C.P. Pipinikas, Y. Jiao, A. Jones, et al., Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer, JAMA Oncol. 1 (4) (2015) 476–485.
[14] Y. Zhang, B. Schöttker, I. Florath, C. Stock, K. Butterbach, B. Holleczek, et al., Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality, Environ. Health Perspect. 124 (1) (2016) 67–74.
[15] R. Zhang, Q. Miao, C. Wang, R. Zhao, W. Li, C.N. Haile, et al., Genome-wide DNA methylation analysis in alcohol dependence, Addict. Biol. 18 (2) (2013) 392–403.
[16] L.P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication, Am. J. Hum. Genet 88 (4) (2011) 450–457.
[17] S. Zeilinger, B. Kühnel, N. Klopp, H. Baurecht, A. Kleinschmidt, C. Gieger, et al., Tobacco smoking leads to extensive genome-wide changes in DNA methylation, PLoS One 8 (5) (2013) e63812.
[18] D. Bönsch, B. Lenz, U. Reulbach, J. Kornhuber, S. Bleich, Homocysteine associated genomic DNA hypermethylation in patients with chronic alcoholism, J. Neural Transm. 111 (12) (2004) 1611–1616.
[19] Alcohol abuse and cigarette smoking are associated with global DNA hypermethylation: results from the German Investigation on Neurobiology in Alcoholism (GINA). 2015;49:97–101.
[20] R.A. Philibert, B. Penaluna, T. White, S. Shires, T. Gunter, J. Liesveld, et al., A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs, Epigenetics 9 (9) (2014) 1212–1219.
[21] P.A. Dugué, X. Wang, L. Baglietto, R. Wilson, B. Lehne, E. Makalic, et al., Alcohol consumption is associated with widespread changes in blood DNA methylation: Analysis of cross - sectional and longitudinal data, Addict. Biol. 26 (1) (2021) e12855.
[22] L.E. Wilson, Z. Xu, S. Harlid, A.J. White, M.A. Troester, D.P. Sandler, et al., Alcohol and DNA methylation: an epigenome-wide association study in blood and normal breast tissue, Am. J. Epidemiol. 188 (6) (2019) 1055–1065.
[23] R. Zhao, R. Zhang, W. Li, Y. Liao, J. Tang, Q. Miao, et al., Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence, Asia -Pac. Psychiatry 5 (1) (2013) 39–50.
[24] L. Tsaprouni, T. Yang, J. Bell, K. Dick, S. Kanoni, J. Nisbet, et al., Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation, Epigenetics 9 (10) (2014) 1382–1396.
[25] F. Guida, T.M. Sandanger, R. Castagné, G. Campanella, S. Polidoro, D. Palli, et al., Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation, Hum. Mol. Genet 24 (8) (2015) 2349–2359.
[26] S. Ambatipudi, C. Cuenin, H. Hernandez-Vargas, A. Ghantous, F. Le Calvez-Kelm, R. Kaaks, et al., Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study, Epigenomics 8 (5) (2016) 599–618.
[27] P. Dugué, C. Jung, J.E. Joo, X. Wang, E. Ming, E. Makalic, et al., Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility, Epigenetics 15 (4) (2020) 358–368.
[28] H.R. Elliott, T. Tillin, W.L. McArdle, K. Ho, A. Duggirala, T.M. Frayling, et al., Differences in smoking associated DNA methylation patterns in South Asians and Europeans, Clin. Epigenetics 6 (1) (2014) 4.
[29] H. Alghanim, W. Wu, B. Mccord, DNA methylation assay based on pyrosequencing for determination of smoking status, Electrophoresis 39 (21) (2018) 2806–2814.
[30] S.C.E. Maas, A. Vidaki, R. Wilson, A. Teumer, F. Liu, J.B.J.Van Meurs, Validated inference of smoking habits from blood with a finite DNA methylation marker set, Eur. J. Epidemiol. 34 (11) (2019) 1055–1074.
[31] R. Philibert, J.A. Mills, M. Dogan, S.R.H. Beach, J.D. Long, AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA, Am. J. Med Genet Part B Neuropsychiatr. Genet 183 (1) (2020) 51–60.
[32] J.D. Chamberlain, S. Nusslé, L. Chapatte, C. Kinnaer, D. Petrovic, S. Pradervand, et al., Blood DNA methylation signatures of lifestyle exposures: tobacco and alcohol consumption, Clin. Epigenetics 14 (1) (2022) 155.
[33] D.L. McCartney, A.J. Stevenson, R.F. Hillary, R.M. Walker, M.L. Bermingham, S. W. Morris, et al., Epigenetic signatures of starting and stopping smoking, EBioMedicine 37 (2018) 214–220.
[34] C. Liu, R.E. Marioni, A.K. Hedman, L. Pfeiffer, P.C. Tsai, L.M. Reynolds, et al., A DNA methylation biomarker of alcohol consumption, Mol. Psychiatry 23 (2) (2018) 422–433.
[35] D.L. McCartney, R.F. Hillary, A.J. Stevenson, S.J. Ritchie, R.M. Walker, Q. Zhang, et al., Epigenetic prediction of complex traits and death, Genome Biol. 19 (1) (2018) 136.
[36] S.C.E. Maas, A. Vidaki, A. Teumer, R. Costeira, R. Wilson, J. van Dongen, et al., Validating biomarkers and models for epigenetic inference of alcohol consumption from blood, Clin. Epigenetics 13 (1) (2021) 198.
[37] M. Hattab, S. Clark, E. van den Oord, Overstimation of the classification accuracy of a biomarker for assessing heavy alcohol use. Mol. Psychiatry 23 (11) (2018) 2114–2115.
[38] P.D. Yousefi, R. Richmond, R. Langdon, A. Ness, C. Liu, D. Levy, et al., Validation and characterisation of a DNA methylation alcohol biomarker across the life course, Clin. Epigenetics 11 (1) (2019) 163.

[39] X. Gao, Y. Zhang, L.P. Breitling, H. Brenner, Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration, Oncotarget 7 (30) (2016) 46878–46889.
[40] M. Stephenson, S. Bollepalli, E. Cazaly, J.E. Salvatore, W.F. Street, Associations of alcohol consumption with epigenome-wide DNA methylation and epigenetic age acceleration: individual-level and co-twin comparison analyses, Alcohol Clin. Exp. Res 45 (2) (2022) 318–328.
[41] J.K. Kresovich, A.M.M. Lopez, E.L. Garval, Z. Xu, A.J. White, P. Dale, et al., Alcohol consumption and methylation-based measures of biological age, J. Gerontol. Ser. A Biol. Sci. Med Sci. 76 (12) (2021) 2107–2111.
[42] Elliott P. Airwave [Data set]. Dementias Platform UK. [Internet]. 2017. Available from: ⟨https://doi.org/10.48532/002000⟩.
[43] P. Elliott, A.C. Vergnaud, D. Singh, D. Neasham, J. Spear, A. Heard, The airwave health monitoring study of police officers and staff in great britain: Rationale, design and methods, Environ. Res 134 (2014) 280–285.
[44] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinforma. 12 (2011) 77.
[45] D.W.J. Hosmer, S. Lemeshow, R.X. Sturdivant. Applied logistic regression, Third edit.,, John Wiley & Sons,, 2013.
[46] W. Venables, B. Ripley, Modern Applied Statistics with S, Springer New York,, Springer US, 2002.
[47] Kassambara A., Mundt F. Factoextra: Extract and visualize the results of multivariate data analyses [Internet]. 2020. Available from: ⟨https://cran.r-project.org/package=factoextra⟩.
[48] Koenker R., Portnoy S., Ng P., Zeileis A., Grosjean P., Ripley B. Package quantreg: Quantile Regression. 2015.
[49] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares De Cal, 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 24 (2016) 65–74.
[50] Alfons A. Package cvTools: Cross-validation tools for regression models. 2015.
[51] Wickham H., Chang W. Package ggplot2: An implementation of the grammar of graphics. 2015.
[52] R. Team, R. Core, A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.
[53] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49 (12) (1996) 1373–1379.
[54] A. Vidaki, B. Planterose Jiménez, B. Poggiali, V. Kalamara, K.J. van der Gaag, S.C. E. Maas, et al., Targeted DNA methylation analysis and prediction of smoking habits in blood based on massively parallel sequencing, Forensic Sci. Int Genet 65 (2023) 102878.
[55] R. Philibert, A quantitative epigenetic approach for the assessment of cigarette consumption, Front Psychol. 6 (2015) 656.
[56] Y. Zhang, I. Florath, K. Saum, H. Brenner, Self-reported smoking, serum cotinine, and blood DNA methylation, Environ. Res 146 (2016) 395–403.
[57] N. Kondratyev, A. Golov, M. Alfimova, T. Lezheiko, V. Golimbet, Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation, Clin. Epigenetics 10 (1) (2018) 130.
[58] N.S. Shenker, P.M. Ueland, S. Polidoro, K. Van Veldhoven, F. Ricceri, R. Brown, et al., DNA methylation as a long-term biomarker of exposure to tobacco smoke, Epidemiology 24 (5) (2013) 712–716.
[59] X. Gao, M. Jia, Y. Zhang, L.P. Breitling, H. Brenner, DNA methylation changes of whole blood cells in response to active smoking exposure in adults: A systematic review of DNA methylation studies, Clin. Epigenetics 7 (2015) 113.
[60] N.S. Shenker, S. Polidoro, K. van Veldhoven, C. Sacerdote, F. Ricceri, M.A. Birrel, et al., Epigenome-wide association study in European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking, Hum. Mol. Genet 22 (5) (2013) 843–851.

DISCUSIÓN GENERAL

5. DISCUSIÓN GENERAL

La genética forense es una disciplina que ha evolucionado en gran medida, no solo adaptándose a las nuevas tecnologías, sino también introduciendo nuevas aplicaciones en su repertorio con el objetivo de ayudar en la resolución de investigaciones criminales. El desarrollo de nuevas aplicaciones para marcadores ya conocidos o el descubrimiento de nuevos biomarcadores ha permitido expandir las capacidades de este ámbito a cotas insospechables, presentándose por delante un futuro desafiante a la par que esperanzador. En este marco de crecimiento y expansión, el estudio de la metilación del ADN ha extendido el alcance de las denominadas *DNA intelligence tools*, incorporando a su colección de técnicas la posibilidad de estimar la edad cronológica de un individuo, identificar tejidos y proporcionar información sobre estilos de vida. Estos descubrimientos y su aplicabilidad han copado el interés de una parte del campo en los últimos años, generándose una gran cantidad de conocimiento que ha permitido el desarrollo de modelos de predicción de considerable interés forense. Los modelos diseñados y las múltiples investigaciones en este campo han permitido construir una telaraña de publicaciones que proporcionan profundidad a esta relativamente nueva rama de la genética forense. A fin de contribuir al crecimiento y desarrollo de estas aplicaciones, durante el desarrollo de esta tesis doctoral, se han planteado diferentes aproximaciones a fin de generar modelos de predicción de la edad individual, modelos de identificación de tejidos y modelos de clasificación en base a estilos de vida.

La predicción de la edad ha sido, y es, la aplicación de la metilación del ADN que mayor interés ha suscitado en el ámbito forense. Desde la publicación del modelo multitejido desarrollado por Horvath (263) y ante su inaplicabilidad, por su alto número de marcadores, a casuística forense, se han generado multitud de modelos que buscan contestar, desde un prisma forense, a la pregunta ¿qué edad cronológica tiene el donante de un vestigio biológico? La genética forense, en esta y otras aplicaciones, se enfrenta a los problemas que se le plantean con multitud de incógnitas, éstas, durante el desarrollo de las metodologías que pretenden resolverlas, deben ser asumidas limitando en cierta medida los modelos desarrollados. La llegada de una muestra desconocida al laboratorio, si nos centramos en la predicción de la edad, nos obliga en ciertas ocasiones a realizar asunciones que pueden condicionar los resultados, como por ejemplo el tejido de origen, el rango de edad al que pertenece, la ancestralidad, los estilos de vida y las enfermedades del donante. Es lógico aceptar la inabarcabilidad de la evaluación de todos estos factores a la hora de generar modelos de predicción y la posibilidad de emplear otras técnicas antes del análisis de la metilación para analizar alguna de estas variables, pero su evaluación tiene siempre un valor importante. Para la predicción de la edad, con la construcción de múltiples modelos a lo largo de los años, se han identificado algunos factores que probablemente no tengan un gran efecto en los marcadores específicos de la edad seleccionados en la generación de los modelos de predicción. Estas asunciones son aceptables ya que se asientan sobre la pluralidad de los conjuntos de individuos analizados, no siendo sesgados para el análisis por sus estilos de vida o enfermedades, representando los modelos, por tanto, parte de la variabilidad observable en la población. La aceptación de estas asunciones

puede ser discutible y es innegable la necesidad de tener en cuenta, si es posible, este tipo de factores al analizar la metilación del ADN, pero hay otros que ciertamente pueden tener un efecto determinante. De buena tinta sabemos el impacto que tiene el tipo de tejido en los patrones de metilación, pero en este ámbito no se han tomado riesgos y se han desarrollado modelos específicos de tejido, algo que comentaremos más adelante. Por otro lado, existe otra variable que debe tenerse en cuenta y afecta al modelo desarrollado, el rango de edad empleado durante la construcción del modelo. Matemáticamente, aunque posible, no es recomendable aplicar un modelo construido en base a un determinado rango de edades a muestras que se encuentran fuera del mismo. El modelo puede extrapolar los valores asumiendo que siguen una tendencia similar al resto de valores observados, pero esto no siempre tiene porque ser así. La necesidad de tener en cuenta esta variable se refuerza con lo observado en el modelo de edad desarrollado por Freire-Aradas et al. para menores de 18 años (281), donde el marcador analizado en el gen *KCNAB3* presenta correlación lineal con la edad en menores, pero los valores de metilación se mantienen más o menos estables a partir de la adolescencia hasta la vejez. A fin de evitar problemas derivados se ha considerado la introducción de menores en diversos modelos de predicción de la edad (195,268,276,280,285). El más reciente de estos modelos, desarrollado por Wozniak et al. (276), cubre un amplio rango de edades (1 a 75 años), pero los infantes representan un pequeño número de la totalidad de individuos analizados (N: 112), estando la representación de 1 a 18 años compuesta por un único individuo por edad. Tanto en este trabajo como en otros, es destacable como el número de muestras de menores es anecdótico, conformando el grueso de los individuos el conjunto a partir de 18 años. Con el objetivo de construir un modelo que abarcase el mayor rango de edades posible se combinaron los datos resultantes de la metodología de EpiTYPER de dos trabajos desarrollados en nuestro laboratorio (271,281), construyéndose un modelo de predicción de la edad para muestras de sangre de individuos de entre 2 a 104 años (**artículo 1**). Dicho modelo, compuestos por siete posiciones CpG (*ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* y cg07082267), con aproximadamente diez muestras por edad (N: 895), presentó errores de ± 3,36 y ± 3,41 años para las metodologías estadísticas que mejores resultados presentaron, regresión cuantil con redes neuronales (QRNN) y regresión cuantil con vectores de soporte (QRSVM), respectivamente. Los marcadores empleados en este modelo han sido utilizados previamente en el desarrollo de un modelo de predicción de la edad para individuos adultos (271), estando algunos de ellos, *ASPA*, *PDE4C*, *FHL2* y cg07082267, correlacionados con la edad en menores (281). El estudio de estos marcadores con un rango de edad ampliado (2 a 104 años) ha proporcionado correlaciones con la edad superiores (correlación de *Spearman* (r_s) > 0,85 y < 0,98) a las observadas con los conjuntos de datos de forma individual (r_s adultos > 0,75 y < 0,94, r_s menores > 0,50 y < 0,85). Si comparamos los modelos generados empleando la regresión cuantil en los tres trabajos, esta mayor correlación no se tradujo en un menor error mediano absoluto ($MAE_{mediana}$) o en un mayor porcentaje de clasificaciones correctas dentro de los intervalos de predicción (%CP±PI). Los resultados obtenidos con este modelo son similares a los observados en la actualidad del campo, cercanos a ± 3 años. Aun así, es interesante destacar el posible potencial de los modelos basados en máquinas de aprendizaje, al presentar

los modelos desarrollados con ellas un mayor %CP±PI en comparación con los modelos de regresión cuantil (QR: 78,77%, QRNN: 81,45% y QRSVM: 79,66%).

El modelo estadístico empleado es un pilar fundamental en el desarrollo de un modelo de predicción. La tónica habitual en el campo es el empleo de modelos lineales (267,268,270,275,280), incorporándose recientemente las máquinas de aprendizaje (180,262,274,277,285), generándose errores únicos e independientes de la edad del donante, algo que no acompaña a los patrones de metilación observados. El estudio de la correlación de la edad y la metilación del ADN ha demostrado, de forma consistente, que la variación en los patrones de metilación entre individuos de la misma edad se va acrecentando con el tiempo (Figura 1, **artículo 1**), lo que repercute en la predicción de la edad, observándose mejores predicciones en individuos jóvenes en comparación con personas ancianas (Figura 2, artículo 1). Por tanto, el uso de metodologías estadísticas que permitan definir errores específicos de rango de edad, como es el caso de la regresión cuantil, presentan una ventaja sobre aquellos que aplican el mismo error a todas las edades. Esta capacidad se ha demostrado en el artículo publicado por Freire-Aradas et al. (271), cuyo modelo ($MAE_{mediana}$: ± 3,07 años) permite predecir correctamente la edad de individuos de 20 años con un error de ± 1,26 años. Este comportamiento de los patrones de metilación discordante entre jóvenes y ancianos se podría considerar, en términos matemáticos, como una distribución heterocedástica (varianza no constante) lo que plantea otra limitación para las metodologías tradicionalmente empleadas en la construcción de modelos de predicción. La regresión cuantil (QR) vuelve a anotarse un tanto ya que, al ser un método no paramétrico, no tiene que asumir ninguna hipótesis en relación con el parámetro analizado, en este caso la metilación del ADN. Esto teóricamente les confiere a los modelos basados en QR una mayor robustez, en comparación con otras aproximaciones matemáticas predominantes en el campo, a la hora de enfrentar a datos externos al modelo o, lo que es lo mismo, muestras dubitadas. Si bien es cierto que el modelo presentado en el **artículo 1** presenta ciertas limitaciones que deben ser abordadas, como la tecnología empleada para el análisis de las muestras, se demuestra la importancia de la construcción de modelos de predicción con un número equilibrado de muestras y la robustez y ventajas de la regresión cuantil para la predicción de la edad individual en el ámbito forense.

Como se mencionó en diversos puntos del texto, el tejido ha sido otro factor clave en el desarrollo de los modelos de predicción de la edad. Teniendo en cuenta las grandes diferencias observadas en los patrones de metilación de diferentes poblaciones celulares y la complejidad de los modelos multitejido (debido a la necesidad de emplear un elevado número de marcadores), se comenzaron a desarrollar modelos de predicción contemplando los diferentes vestigios biológicos de interés forense. En los primeros años, la sangre monopolizó el interés del campo, desarrollándose múltiples modelos en base a este tejido. De forma simultánea, aunque en menor medida, se construyeron modelos en base a muestras de saliva e hisopo bucal (195,275,276,289–293), vestigios procedentes de la misma cavidad, pero compuestos por poblaciones celulares diferentes. La presencia de distintas poblaciones y la coexistencia de las muestras en la misma cavidad se ha demostrado que genera un rango de proporciones celulares muy variable en cada muestra (195), incrementando la complejidad del estudio de estos tejidos.

La mayoría de los modelos generados han tratado estos tejidos de forma independiente, generando modelos solo con muestras de saliva (289,290) y modelos solo con muestras de hisopo (195,276,291). Pero con el objetivo de enfrentarnos a muestras forenses desconocidas es necesario plantear otras alternativas para paliar esta variabilidad celular. Por tanto, se han evaluado dos formas de afrontar el análisis de las muestras procedentes de la cavidad oral: desarrollar un modelo de identificación de tejidos para intentar discriminar entre muestras de saliva y de hisopo bucal y, por otro lado, desarrollar un modelo de predicción de la edad combinando ambos tejidos (**artículo 2**). En las situaciones en las que solo se disponga de modelos de predicción de la edad para saliva y mucosa oral de manera independiente sería conveniente intentar identificar el tejido de la muestra desconocida para así aplicar el modelo correspondiente. Actualmente, los modelos de identificación de tejido más destacables abarcan uno de estos dos tejidos, principalmente saliva (204–206), pero no ambos simultáneamente, lo que podría no representar toda la variabilidad asociada a estas muestras, siendo posible la ocurrencia de falsos positivos asociados a otros tejidos que comparten un porcentaje de estas poblaciones celulares, como por ejemplo la presencia de células epiteliales en fluido vaginal. Con esto en mente, empleando dos marcadores específicos de tejido (situados en los genes *HUNK* y *RUNX1*), se desarrolló un modelo de regresión logística que permite identificar saliva e hisopo bucal con un porcentaje de clasificaciones correctas del 88,59%. Aunque este porcentaje es elevado, la complejidad y el amplio rango de proporciones celulares observadas en estas muestras hace complicada su clasificación, complejidad que aumenta al enfrentarse a muestras mezcla de saliva e hisopo, como puede ser durante el análisis de colillas de cigarrillos encontradas en la escena de un crimen. A su vez, la selección de estos marcadores se realizó pensando solamente en la distinción de saliva y mucosa oral, por lo que dichos marcadores solo se han evaluado con muestras de sangre, saliva e hisopo, siendo muy limitada su aplicabilidad individual en el análisis de muestras dubitadas de origen desconocido. Por tanto, se consideró la construcción de modelos de predicción de la edad combinando muestra de saliva e hisopo bucal, una mejor estrategia para afrontar el análisis de muestras o mezclas procedentes de la cavidad oral. Se construyó un modelo compuesto por siete posiciones CpG (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* y *EDARADD*) obteniéndose un $MAE_{mediana}$ de ± 3,54 años y un porcentaje de clasificaciones correctas de 76,08%. Este trabajo presenta un error similar al de otros modelos previamente publicados, empleando aproximaciones estadística diferentes y tratando de forma independiente los tejidos evaluados (276,290,292,293). El modelo de predicción de la edad presentado en el **artículo 2**, sigue una estrategia semejante a la presentada por Jung et al. (275), cuyo modelo es capaz de predecir la edad, no solo de muestras de saliva e hisopo bucal, sino también de sangre, presentando un error de ± 3,55 años. Estos resultados refuerzan el peso de la composición celular en la predicción de la edad, presentando los tres tejidos células leucocitarias y, por tanto, permitiendo la identificación de marcadores correlacionados con la edad, en mayor o en menor medida, en los tres tejidos, como por ejemplo *ELOVL2*, *PDE4C* y *EDARADD*, genes habitualmente asociados a modelos de sangre. Teniendo esto en cuenta, se intentó evaluar, empleando la información de tejido, si la procedencia de las muestras podía ser una variable de peso que mejorase el error de predicción obtenido por el modelo de edad. La introducción del tejido de

origen como covariable, no produjo una mejoría en el rendimiento del modelo. Esto podría deberse a que la variabilidad dentro y entre los tejidos no puede ser explicada por un factor tan general, observándose en la Figura 3 del **artículo 2** como las muestras de saliva e hisopo bucal se solapan parcial o mayoritariamente en todos los marcadores de edad empleados en el modelo. Si comparamos este análisis con el planteado por Eipel et al. (195), que observó una mejoría en la predicción de la edad al introducir una covariable que aporta información de la composición celular de la muestra, podemos concluir que el tejido por sí mismo no es lo suficientemente informativo, dando de nuevo un gran peso a las proporciones celulares observadas en las muestras procedentes de la cavidad oral, efecto extensible al resto de tejidos de interés forense. Este acercamiento podría ser un planteamiento que ofreciese una mejoría en los futuros modelos de predicción, pero aún es necesario realizar más estudios para poder afirmarlo. Esta información adicional podría ser clave en la construcción de modelos de predicción de la edad multitejido con un bajo número de marcadores, algo que aún debe ser explorado en detalle.

Ante la inminente aplicación rutinaria de los modelos de predicción de la edad en casuística forense se han comenzado a plantear y analizar otros inconvenientes asociados a la metilación del ADN. Uno de los más relevantes, que podría en cierta medida limitar su uso, es el tipo de metodología empleada para desarrollar el modelo. La intercambiabilidad de los datos de metilación entre plataformas ha sido abordada (249), pero se ha demostrado factible principalmente entre metodologías cuantitativas (pirosecuenciación, EpiTYPER y MiSeq®) siempre y cuando las diferencias en los niveles de metilación de las plataformas evaluadas no superen límites de uniformidad. La estabilidad de detección, por parte de estas tecnologías, de los patrones de este biomarcador, se ha demostrado que es extensible a los paneles *BeadChip* de Illumina (289), un factor clave y de gran importancia ya que los estudios de descubrimiento de marcadores se llevan a cabo, predominantemente, empleado dichos paneles. Si bien es cierto que estos resultados son interesantes, debemos ser cuidadosos ya que estas asunciones podrían ser marcador-específicas, observándose falta de uniformidad entre plataformas para uno de los marcadores analizados (249). Este panorama cambia cuando comenzamos a trabajar con tecnologías semicuantitativas, como SNaPshot, que emplea en la detección de fluorocromos que presentan diferentes intensidades de señal. Este hecho juega un papel importante que provoca que esta tecnología presente niveles de metilación discordantes en relación con las técnicas cuantitativas (249), pero también entre plataformas de análisis de SNaPshot (250). Al identificarse estas discordancias se han propuesto diversos acercamientos con el objetivo de facilitar el uso de los modelos independientemente de la tecnología de análisis empleada (211,246–248). Gracias a estos trabajos se han planteado diferentes aproximaciones que permiten corregir los valores detectados o el modelo desarrollado, ampliando su aplicabilidad. Con esto en mente y teniendo en cuenta que el modelo de predicción de la edad presentado en el **artículo 2** ha sido desarrollado empleando SNaPshot, analizando las muestras en un secuenciador capilar (ABI3130xl) que ha sido recientemente descontinuado, se presenta la necesidad de adaptar dicho modelo a un nuevo equipo. Siguiendo la estrategia planteada por So et al. (248), se desarrollaron modelos de transformación para los niveles de metilación detectados en el ABI3500 para un total de 114 muestras empleadas en la construcción del

modelo de predicción de la edad (ABI3130xl) desarrollado. Las diferencias observadas entre los patrones de metilación no eran solo dependientes de la plataforma, el marcador analizado también jugó un papel importante observándose diferencias significativas entre marcadores (p-valor de $2{,}20^{-16}$), por tanto, se desarrollaron modelos de transformación específicos para cada marcador. Los resultados presentados en el **Anexo I**, permiten la correcta predicción de muestras analizadas en el ABI3500 empleando el modelo de saliva e hisopo bucal desarrollado en base a otro secuenciador capilar, el ABI3130xl. La extensión de la aplicabilidad de este modelo permite alargar su vida útil limitada por la metodología empleada, siendo necesario en el futuro el diseño de modelos que no estén condicionados a una o dos plataformas específicas. Otro posible acercamiento, que no ha podido ser explorado, es la construcción o transformación de este modelo a una tecnología cuantitativa que ha ganado protagonismo en los últimos años, MiSeq®. Este paso permitiría extender en mayor medida la viabilidad, aplicabilidad y longevidad del modelo desarrollado.

Con el desarrollo de modelos de predicción que proporcionan errores asumibles en estos tejidos, se ha explorado el estudio de otros vestigios de interés forense, como por ejemplo los restos cadavéricos. Entre ellos se han construido modelos de predicción de la edad para dientes, hueso, pelo y uñas (276,295,305,307,309), pero existen otros que aún no han sido explorados, como por ejemplo el cartílago. En lo referente al análisis de la metilación del ADN, la predicción de la edad para este tejido se ha intentado llevar a cabo empleando un modelo desarrollado para huesos (276), pero los errores de predicción obtenidos fueron muy elevados (± 25,8 años). Por tanto, se ha desarrollado, hasta donde sabemos, el primer modelo de predicción de la edad específico de cartílago basado en metilación del ADN (**artículo 3**). El modelo de regresión cuantil construido está compuesto por tres posiciones CpG (comprendidas en los genes *FHL2*, *TRIM59* y *KLF14*) y presenta un error de predicción de ± 4,41 años (**artículo 3**). Si bien es cierto que el error es ligeramente superior al presentado por los modelos de predicción de la edad actuales, no se aleja mucho del límite que en los últimos años parece haberse alcanzado (aproximadamente ± 3 años). Este resultado podría ser justificado por la sobreestimación de la edad predicha observada en las muestras de individuos jóvenes y la infraestimación observada en muestras de personas de mayor edad, como se puede observar en la Figura 1 del **artículo 3**. Por otro lado, este efecto también podría estar justificado por el limitado número de muestras (N: 109), condicionado por la disponibilidad del tipo de tejido analizado, pudiendo ser relevante la ampliación del conjunto de entrenamiento. Por tanto, aun siendo un primer acercamiento que presenta buenos resultados es recomendable evaluar en mayor detalle este tejido. A su vez, gracias a la disponibilidad de muestra analizadas en el proyecto europeo VISAGE (276) y teniendo en cuenta el gran impacto que puede ocasionar en los errores de predicción el tejido de origen, se desarrolló un modelo de identificación de tejido para muestras de sangre, hisopo bucal, hueso y cartílago (**artículo 3**). El modelo de regresión logística multinomial generado está compuesto por ocho posiciones CpG (comprendidas en los genes *EDARADD*, *TRIM59*, *ELOVL2*, *MIR29B2CHG*, *PDE4C*, *ASPA*, *FHL2* y *KLF14*) y presenta un porcentaje de clasificaciones correctas de 98,72%. Este modelo de identificación de tejido presenta casi una clasificación perfecta, si nos fijamos en la Figura 2 del **artículo 3**,

observándose cómo un pequeño número de muestras de dos tejidos (hisopo bucal y hueso) solapan ligeramente, estando la mayoría de las muestras completamente separadas. Esta superposición podría deberse a la existencia de una presencia minoritaria de leucocitos en hisopo bucal y hueso, presentando cierta tendencia ambos tejidos hacia la sangre. Esta baja proporción de células sanguíneas puede justificarse del siguiente modo: los huesos requieren de un suministro de sangre, fluido que viaja a través del periostio (parte del hueso) hasta la médula ósea interior, y la muestra que definimos como hisopo bucal es el raspado de las paredes de la cavidad oral, estando en contacto con la saliva, un tejido que presenta una proporción mayoritaria de leucocitos. Por otro lado, es interesante destacar que las posiciones CpG empleadas en el modelo de identificación de tejido se encuentran en genes comúnmente asociados con predicción de la edad. Si bien es cierto que los marcadores específicos de tejido no deben estar correlacionados con la edad, la cercanía entre posiciones CpG de distintas aplicaciones podría facilitar el diseño experimental de modelos multitejido en plataformas de secuenciación masiva en paralelo (MiSeq®), permitiendo tanto la identificación de tejido como la predicción de la edad cronológica del donante de la muestra. Modelos de identificación de tejido como éste, con un alto porcentaje de clasificaciones correctas, tras una correcta optimización y validación, podrían implementarse en casuística forense y sustituir a alguno de los análisis químicos e inmunológicos que se emplean actualmente, ya que dichos métodos son incapaces de identificar la totalidad de tejidos de interés forense.

La identificación de tejidos y la predicción de la edad en base a metilación de ADN ha avanzado a pasos agigantados en los últimos años, presentando resultados muy prometedores y, en el caso de los modelos de edad, han comenzado a ser aplicados en casuística forense. Centrándonos en la predicción de la edad es importante destacar que en los últimos años parece haberse alcanzado una limitación temporal, observándose un error mínimo de predicción cercano a ± 3 años, independientemente del tejido analizado. Esto plantea una pregunta, ¿cuáles son los siguientes pasos a tomar para obtener errores de predicción más bajos? La respuesta quizás esté en las proporciones de las poblaciones celulares que pueden encontrarse en los distintos tejidos analizados o quizás sea hora de seguir los pasos de otras aplicaciones, como la ancestralidad, y construir, gracias a las tecnologías de MPS, modelos con un mayor número de marcadores que permitan simultáneamente identificar y predecir la edad de varios tejidos.

A lo largo de los años, otras aplicaciones de interés tanto forense como clínico han sido evaluadas empleando el análisis de la metilación del ADN. La comprensión del funcionamiento de este biomarcador ha permitido identificar el importante efecto producido por el ambiente a lo largo del epigenoma. Dentro del ambiente, los factores más estudiados son los efectos de los estilos de vida en los patrones de metilación, evaluándose las alteraciones en la expresión génica y su asociación con distintas enfermedades. A nivel forense, estos descubrimientos han abierto una nueva rama que puede proporcionar información sobre el donante de un vestigio biológico, comenzando a verse los frutos de este florecimiento en los últimos años. Principalmente, el ámbito forense se ha centrado en la predicción del consumo de dos sustancias perjudiciales, el tabaco y el alcohol. La primera ha mostrado resultados llamativos desde un primer momento (336), e incluso se han alcanzado, en modelos de clasificación de no fumadores *vs* fumadores,

AUCs de 0,97 empleando un solo marcador (348). El consumo de alcohol, por otro lado, se ha tornado más desafiante, siendo un estilo de vida con un efecto más complejo sobre los patrones de metilación que el tabaco, existiendo diversos factores que tienen un peso más significativo en su evaluación, como, por ejemplo, la información de consumo aportada por los individuos, las categorías definidas, el efecto general de este estilo de vida en el epigenoma y el sexo como una variable de peso (366,369–371). Aún con todos estos inconvenientes, en los últimos años se han generado modelos forenses de clasificación tanto para el consumo de tabaco como para el consumo de alcohol. Pero muchos de los modelos desarrollados han obviado a una importante porción de la población, realizando solamente una clasificación entre los grupos extremos, no fumadores *vs* fumadores y no bebedores *vs* bebedores altos, pero ¿qué pasa con las categorías intermedias? ¿qué pasa con los ex consumidores o, en el caso de alcohol con aquellos que presentan un consumo más moderado? La importancia de la representación de estos grupos radica en la necesidad de representar a la población general en los modelos que a ella quieren aplicarse, pero debemos ser cuidadosos con los patrones que presentan estos grupos. Las alteraciones en los patrones de metilación por tabaco y alcohol se ha demostrados que son reversibles en algunas posiciones, recuperando, sobre todo en tabaco, valores cercanos a los de no consumidores (341–344,364). Esta restauración de la metilación supone un punto de inflexión y pone sobre la mesa la complejidad de dicha aplicación. La clasificación de una categoría que se encuentra transitando entre los extremos en base a factores asociados al consumo, como tiempo consumido, intensidad o tiempo desde cese, se convierte en un desafío. A su vez, en el caso de alcohol, se observa una mayor diversidad de grados de consumo y unos efectos menos marcados sobre el epigenoma, lo que dificulta la definición de las categorías que componen a los bebedores, creando grupos intermedios que engloban una gran proporción de consumo o segmentaciones de éste que buscan estrechar más las categorías analizadas. Ante esta diversidad de categorías, a fin de evaluarlas todas, se ha planteado el uso de modelos multinomiales, pero los resultados obtenidos para las categorías intermedias están lejos de ser aplicables (346,347). Por tanto, deben explorarse otros horizontes a fin de construir modelos aplicables a la población general que proporcionen información relevante con respecto a estos estilos de vida. Con esto en mente, se han desarrollado modelos de clasificación para consumo de tabaco y alcohol de uso forense (**artículo 4**). En nuestro trabajo también se llevó a cabo una aproximación estadística multinomial con modelos de regresión logística, pero los resultados presentaron las tendencias ya observadas en la bibliografía. Este acercamiento no fue infructuoso, permitiéndonos evaluar el comportamiento del grupo intermedio. Si nos fijamos en la Tabla 3 del **artículo 4** podemos observar como el grupo intermedio se clasifica mejor en los modelos de tabaco en comparación con los de alcohol, pero conviene que nos fijemos donde se clasifica "mal". En el caso de tabaco, el grupo mayoritario se clasifica correctamente como exfumadores (≈ 50%), pero se observa cómo más de un 35% de los individuos de este grupo se clasifican incorrectamente como no fumadores, algo que podría ser esperable teniendo en cuenta la reversibilidad observada en marcadores asociados con el consumo de tabaco. Esta observación es más llamativa en el caso del alcohol, donde de forma consistente, la mayoría de los bebedores moderados son clasificados como no bebedores (≈ 50%). Teniendo estos resultados en cuenta y ante la complejidad asociada a la correcta clasificación de estas

categorías, se decidió llevar a cabo otra aproximación, la agrupación de categorías menos informativas a nivel forense con el objetivo de predecir correctamente aquellas más individualizantes. Teniendo en cuenta las estadísticas ya comentadas, los grupos menos representados en la población y que, por tanto, más podrían acotar el número de sospechosos en una identificación son los grupos de fumadores y bebedores altos. Por tanto, se desarrollaron modelos logísticos binomiales enfrentando, por un lado, no fumadores + exfumadores *vs* fumadores y, por otro lado, no bebedores + bebedores moderados *vs* bebedores altos, presentando dichos modelos AUCs de 0,87 y 0,80, sensibilidad de 0,79 y 0,81, especificidad de 0,90 y 0,71 y porcentajes de clasificaciones correctas de 86,49% y 74,26%, respectivamente. La unión de estas categorías se ve reforzada por la bibliografía, observándose una mayor diferenciación entre exfumadores *vs* fumadores que entre no fumadores *vs* exfumadores (AUC medio de 0,91 y 0,73 , respectivamente) (346). Por otro lado, en el caso de alcohol, al construir un modelo agrupando consumidores (bebedores altos + bebedores moderados) frente a no consumidores se obtuvo un AUC de 0,64 (340). Con estos resultados y los observados en el **artículo 4**, se puede concluir que, si los modelos multinomiales no permiten una correcta separación de las categorías evaluadas para estos estilos de vida, es recomendable agrupar la categoría intermedia con los no consumidores. Este acercamiento fue planteado previamente tanto para la evaluación del consumo de tabaco (347) como para el consumo de alcohol (369), obteniéndose con un mayor número de marcadores, AUCs o sensibilidades inferiores a las obtenidas en el modelo desarrollado en el **artículo 4**. Si bien es cierto que los modelos de clasificación desarrollados presentan mejores resultados en comparación con los otros modelos, aún queda un largo camino por recorrer. Con el objetivo de mejorar los modelos actuales sería recomendable abordar factores asociados a las dinámicas de consumo tanto de tabaco como de alcohol. En el caso de tabaco ya se han realizado los primeros acercamientos hacia la predicción del número de cigarrillos consumidos por día, años como fumador o tiempo desde cese de consumo (336,347,351), variables que de ser usada en los modelos de clasificación de estado de fumadores podrían ser de gran utilidad para aportar información sobre los exfumadores y fumadores.

Más allá de la identificación de patrones de consumo, cabe destacar que los efectos de estos estilos de vida en la metilación del ADN se han asociado con alteraciones relacionadas con patrones de aceleración del envejecimiento biológico (252,255,256). Esta asociación y la existencia de modelos de predicción de la edad, tanto cronológica como biológica, refuerza la necesidad de tener en cuenta estas variables como posibles cofactores en la predicción de la edad y, por tanto, se ha llevado a cabo una evaluación de este efecto (**artículo 4**). Para ello, se construyó un modelo de predicción de la edad basado en los marcadores empleados en una publicación previa (271) y se evaluaron los residuos de los modelos para los grupos previamente definidos para la clasificación de tabaco y alcohol. En ambos casos no se observaron diferencias significativas (p-valor de 0,78 y 0,10 grupos evaluados en tabaco y alcohol; respectivamente) entre las edades predichas de las categorías evaluadas, por lo que los marcadores analizados no parecen estar influenciados por dichos estilos de vida. Es importante destacar que estos

resultados son específicos de los marcadores evaluados, no siendo extrapolables a otros y siendo su evaluación recomendable siempre que sea posible.

El análisis de la metilación del ADN es, como hemos podido observar a lo largo de este texto, un campo complejo, pero gracias al cuidadoso estudio de científicos de todo el mundo se ha construido un ecosistema de conocimiento que ha permitido su aplicación al ámbito forense. Mientras que el impacto de los modelos de predicción de la edad es cada vez más una realidad en casuística forense, nuevas aplicaciones comienzan a abrirse paso ágilmente gracias a los conocimientos previamente establecidos. Las posibilidades de esta marca epigenética en el ámbito clínico y forense seguirán siendo investigadas en los próximos años, la complejidad de su estudio seguirá manifestándose y dificultará el camino, pero bien haríamos en recordar esta cita del libro El problema de los tres cuerpos de Liu Cixin: "*La curiosidad es una de las cualidades más valiosas del ser humano, ya que nos impulsa a buscar respuestas a las preguntas más difíciles*".

CONCLUSIONES

6. CONCLUSIONES

6.1 Conclusiones sobre la estimación de la edad empleando marcadores epigenéticos

i. Se ha desarrollado con éxito un modelo de regresión cuantil compuesto por siete posiciones CpG (*ELOVL2*, *ASPA*, *PDE4C*, *FHL2*, *CCDC102B*, *MIR29B2CHG* y cg07082267) para la predicción de la edad en base a muestras de sangre. Dicho modelo abarca el mayor rango de edades publicado hasta la fecha (2 a 104 años), presentando errores en concordancia con los resultados esperados en la actualidad del campo ($MAE_{mediana}$: ± 3,36 años y ± 3,41 años para QRNN y QRSVM, respectivamente).

ii. Se ha desarrollado con éxito un modelo de regresión cuantil compuesto por siete posiciones CpG (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* y *EDARADD*) para la predicción de la edad en base a muestras procedentes de la cavidad oral, combinando muestras de saliva e hisopo bucal. Dicho modelo permite abordar vestigios que presenten mezcla de dichos tejidos obteniéndose un $MAE_{mediana}$ de ± 3,54 años.

iii. Se evaluó el efecto del tejido como covariable del modelo de predicción de la edad para muestras de saliva e hisopo bucal, no observándose una mejoría en las predicciones obtenidas. Por tanto, se descartó como variable informativa para dicho modelo.

iv. Teniendo en cuenta el contexto forense, se evaluó la sensibilidad del modelo de predicción de la edad para muestras de saliva e hisopo bucal, observándose resultados consistentes para los patrones de metilación obtenidos empleando un mínimo de 10 ng de ADN genómico de partida. Estos resultados son consistentes con los observados en modelos publicados previamente.

v. La robustez del modelo de predicción para saliva e hisopo bucal desarrollado fue evaluada, identificándose los marcadores cuya pérdida produce una mayor repercusión en el error obtenido por el modelo. Ante muestras degradadas o con baja cantidad de ADN se recomienda el análisis por duplicado de las muestras a fin de asegurar que dichas condiciones no afectan a los patrones de metilación obtenidos, siendo factible el uso de un menor número de marcadores si el error del modelo se mantiene estable.

vi. Se ha desarrollado con éxito un modelo de regresión cuantil compuesto por tres posiciones CpG (*FHL2*, *TRIM59* y *KLF14*) para la predicción de la edad en base a muestras de cartílago. Dicho modelo es el primer modelo forense que aborda este tejido presentando un $MAE_{mediana}$ de ± 4,41 años.

vii. El uso de modelos de regresión cuantil ha demostrado ser una aproximación en consonancia con el comportamiento de los patrones de metilación observados en los marcadores correlacionados con la edad. Esta metodología estadística permite generar errores específicos para cada edad proporcionando mejores predicciones para los rangos de edad que presentan menor variabilidad.

6.2 Conclusiones sobre la identificación de tejido empleando marcadores epigenéticos

i. Se ha desarrollado con éxito un modelo de regresión logística compuesto por dos posiciones CpG (*HUNK* y *RUNX1*) para la identificación de tejido que permite diferenciar saliva de hisopo bucal con un porcentaje de clasificaciones correctas del 88,59%. Este modelo está limitado a los tejidos evaluados no siendo recomendable su uso en muestras desconocidas que puedan presentar tejidos diferentes a los obtenidos a partir de la cavidad oral.

ii. Se ha desarrollado con éxito un modelo de regresión logística multinomial compuesto por ocho posiciones CpG (*EDARADD*, *TRIM59*, *ELOVL2*, *MIR29B2CHG*, *PDE4C*, *ASPA*, *FHL2* y *KLF14*) para la identificación de tejido que permite diferenciar sangre, hisopo bucal, hueso y cartílago con un porcentaje de clasificaciones correctas del 98,72%.

iii. El uso de posiciones CpG asociadas con genes correlacionados con la edad en la identificación de tejidos permite, gracias a las tecnologías de secuenciación masiva en paralelo, la identificación de patrones de metilación simultáneos para predicción de la edad e identificación de tejidos.

6.3 Conclusiones sobre la inferencia de consumo de tabaco y alcohol empleando marcadores epigenéticos

i. Se ha desarrollado con éxito un modelo de regresión logística binomial compuesto por dos posiciones CpG (*AHRR* y cg01940273) para la inferencia de consumo de tabaco, permitiendo clasificar entre "no fumadores + exfumadores" *versus* "fumadores" con un porcentaje de clasificaciones correctas del 86,49%.

ii. Se ha desarrollado con éxito un modelo de regresión logística binomial compuesto por tres posiciones CpG (*SLC7A11*, cg0886875 y *MIR4435–2HG*) para la inferencia de consumo de alcohol, permitiendo clasificar entre "no bebedores + bebedores moderados" *versus* "bebedores altos" con un porcentaje de clasificaciones correctas del 74,26%.

iii. El consumo de tabaco y alcohol no ha mostrado influencia en la predicción de la edad para los marcadores *ELOVL2*, *ASPA*, *FHL2*, *CCDC102B* y cg07082267 (p-valor de 0,78 y 0,10 grupos evaluados en tabaco y alcohol, respectivamente).

BIBLIOGRAFÍA

7. BIBLIOGRAFÍA

1. Sala N, Arsuaga JL, Pantoja-Pérez A, Pablos A, Martínez I, Quam RM, et al. Lethal interpersonal violence in the middle pleistocene. PLoS One. 2015;10(5):e0126589.
2. Trigger B, Kemp B, O'Connor D, Lloyd A. Ancient egypt: A social history. Cambridge University Press; 1983.
3. Wecht CH. The history of legal medicine. J Am Acad Psychiatry Law. 2005;33(2):245–51.
4. Oliver JR. Legal medicine in Europe and America. Am Bar Assoc J. 1932;18:405–11.
5. Spitz WU, Diaz FJ, Fisher RS. Spitz and Fisher's Medicolegal investigation of death: Guidelines for the application of pathology to crime investigation. 5th ed. Charles C. Thomas Publisher; 2020.
6. Fodéré F-E. Traité de Médecine Légale. Imprimerie de Mame, editor. París; 1768.
7. Tilstone W, Savage K, Clark L. Forensic Science: an encyclopedia of history, methods and tecniques. Bloomsbury Academic; 2006.
8. Dunn LC. A short history of genetics: The development of some of the main lines of thought, 1864-1939. J Hist Biol. 1993;26(1):158–9.
9. Landsteiner K. Über agglutinationserscheinungen normalen menschlichen Blutes. Wien Klin Wochenscher. 1901;14:1132–4.
10. Tan SY, Graham C. Karl Landsteiner (1868-1943): Originator of AB0 blood classification. Singapore Med J. 2013;54(5):243–4.
11. Goodwin W, Linacre A, Hadi S. An introduction to Forensic Genetics. Vol. 53, Journal of Chemical Information and Modeling. 2011. 214 p.
12. Dahm R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet. 2008;122(6):565–81.
13. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal yypes: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. J Exp Med. 1944;79(2):137–58.
14. Franklin R, Gosling R. Molecular configuration in sodium thymonucleate. Nature. 1953;171:740–1.
15. Watson J, Crick F. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature. 1953;171:737–8.
16. Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol. 1975;98(3):503–17.
17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74(12):5463–7.
18. Ford EB. Polymorphism and taxonomy. Heredity (Edinb). 1955;9(2):255–64.
19. Wyman AR, White R. A highly polymorphic locus in human DNA. Proc Natl Acad Sci U S A. 1980;77(11):6754–8.
20. Jeffreys AJ, Wilson V, Thein SL. Individual-specific "fingerprints" of human DNA. Nature. 1985;316(6023):76–9.
21. Weller P, Jeffreys AJ, Wilson V, Blanchetot A. Organization of the human myoglobin gene. EMBO J. 1984;3(2):439–46.
22. Gill P, Jeffreys AJ, Werrett DJ. Forensic application of DNA fingerprints. Nature. 1985;318(6046):577–9.
23. Wong Z, Wilson V, Patel I, Povey S, Jeffreys AJ. Characterization of a panel of highly variable minisatellites cloned from human DNA. Ann Hum Genet. 1987;51(4):269–88.
24. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51(10):263–73.
25. Crespillo Márquez MC, Barrio Caballero PA. Genética Forense. Del laboratorio a los tribunales. Díaz de Santos; 2019.
26. Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet. 1989;44(3):388–96.
27. Hagelberg E, Gray IC, Jeffreys AJ. Identification of the skeletal remains of a murder victim by DNA analysis. Nature. 1991;352:427.
28. Jeffreys A, J. Allen M, Hagelberg E, Sonnberg A. Identification of the skeletal remains of Mengele, Josef by DNA analysis. Vol. 56, Forensic science international. 1992. 65–76 p.
29. Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet. 1991;49(4):746–56.
30. Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med. 1997;110:125–33.
31. Edelman J, Hering S, Michael M, Lessig R, Deischel D, Meier-Sundhausen G, et al. 16 X-chromosome STR loci frequency data from a German population. Forensic Sci Int. 2001;124(2–3):215–8.
32. Carracedo A, Lareu M V. Development of new STRs for forensic casework: Criteria for selection, sequencing & population data and forensic validation. Proceedings—the Ninth Int Symp Hum Identif. 1998;89–107.

33. Lareu MV, del Carmen Pestoni M, Barros F, Salas A, Carracedo A. Sequence variation of a hypervariable short tandem repeat at the D12S391 locus. Gene. 1996;182(1):151–3.
34. Lareu M V, Barral S, Salas A, Pestoni C, Carracedo A. Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus. Int J Legal Med. 1998;111(5):244–7.
35. Kimpton C, Fisher D, Watson S, Adams M, Urquhart A, Lygo J, et al. Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int J Legal Med. 1994;106(6):302–11.
36. The ethics of catching criminals using their family's DNA. Nature. 2018;557(7703):5.
37. Sharma V, Wurmbach E. Systematic evaluation of the Precision ID GlobalFiler™ NGS STR panel v2 using single-source samples of various quantity and quality and mixed DNA samples. Forensic Sci Int Genet. 2024;69:102995.
38. Tillmar A, Sturk-Andreaggi K, Daniels-Higginbotham J, Thomas JT, Marshall C. The FORCE panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications. Genes (Basel). 2021;12(12):1968.
39. Collins FS, Fink L. The human genome project. Alcohol Health Res World. 1995;19(3):190–5.
40. Hendrick PW. Genetics of populations. Jones & Bartlett Publishers; 2005. 104–105 p.
41. Fan H, Chu JY. A brief review of Short Tandem Repeat mutation. Genomics, Proteomics Bioinforma. 2007;5(1):7–14.
42. Jobling MA, Hurles ME, Tyler-Smith C. Human evolutionary genetics: Origins, peoples & disease. Vol. 47, Journal of Human Evolution. Garland Science; 2004.
43. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014;24(11):1894–904.
44. Zajac P, Oberg C, Ahmadian A. Analysis of short tandem repeats by parallel DNA threading. PLoS One. 2012;4(11):e7823.
45. Jefatura del Estado. Ley Orgánica 10/2007, de 8 de octubre, reguladora de la base de datos policial sobre identificadores obtenidos a partir del ADN. España: BOE; 2007 p. 40969–72.
46. Gharesouran J, Hosseinzadeh H, Ghafouri-Fard S, Taheri M, Rezazadeh M. STRs: Ancient architectures of the genome beyond the sequence. J Mol Neurosci. 2021;71(12):2441–55.
47. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
48. Urquhart A, Kimpton CP, Downes TJ, Gill P. Variation in short tandem repeat sequences -- a survey of twelve microsatellite loci for use as forensic identification markers. Int J Legal Med. 1994;107(1):13–20.
49. Phillips C, Fernandez-Formoso L, Gelabert-Besada M, García-Magariños M, Amigo J, Carracedo Á, et al. Global population variability in Qiagen Investigator HDplex STRs. Forensic Sci Int Genet. 2014;8(1):36–43.
50. Brookes AJ. The essence of SNPs. Gene. 1999;234(2):177–86.
51. Miller RD, Phillips MS, Jo I, Donaldson MA, Studebaker JF, Addleman N, et al. High-density single-nucleotide polymorphism maps of the human genome. Genomics. 2005;86(2):117–26.
52. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
53. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998;280(5366):1077–82.
54. Budowle B, Van Daal A. Forensically relevant SNP classes. Biotechniques. 2008;44(5):603–10.
55. Gill P. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Vol. 114, International journal of legal medicine. 2001. 204–210 p.
56. Krawczak M. Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis. 1999;20(8):1676–81.
57. Budowle B. SNP typing strategies. Vol. 146 Suppl, Forensic science international. 2005. S139-42 p.
58. R. Hughes-Stamm S, J. Ashton K, van Daal A. Assessment of DNA degradation and the genotyping success of highly degraded samples. Int J Legal Med. 2011;125(3):341–8.
59. Amorim A, Pereira L. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic Sci Int. 2005;150(1):17–21.
60. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. Human population genetic structure and inference of group membership. Am J Hum Genet. 2003;72(3):578–89.
61. Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Álvarez-Dios J, et al. Ancestry analysis in the 11-M Madrid bomb attack investigation. PLoS One. 2009;4(8):e6583.
62. Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, et al. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int Genet. 2018;35:123–35.
63. White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet. 2021;53(1):45–53.
64. Pereira V, Freire-Aradas A, Ballard D, Børsting C, Diez V, Pruszkowska-Przybylska P, et al. Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel. Forensic Sci Int

Genet. 2019;42:260–7.
65. de la Puente M, Ruiz-Ramírez J, Ambroa-Conde A, Xavier C, Pardo-Seco J, Álvarez-Dios J, et al. Development and evaluation of the ancestry informative marker panel of the visage basic tool. Genes (Basel). 2021;12(8):1284.
66. Ruiz-Ramírez J, de la Puente M, Xavier C, Ambroa-Conde A, Álvarez-Dios J, Freire-Aradas A, et al. Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry. Forensic Sci Int Genet. 2023;64:102853.
67. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001;29:229–32.
68. Ceppelini R, Curtoni E, Mattiuz P, Miggiano V, Scudeller G, Serra A. Genetics of leukocyte antigens: A family study of segregation and linkage. Histocompat Test. 1967;149–87.
69. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
70. Kidd KK, Pakstis AJ, Speed WC, Lagace R, Chang J, Wootton S, et al. Microhaplotype loci are a powerful new type of forensic marker. Forensic Sci Int Genet Suppl Ser. 2013;4:e123–4.
71. Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. Forensic Sci Int Genet. 2019;38:54–69.
72. de la Puente M, Ruiz-Ramírez J, Ambroa-Conde A, Xavier C, Amigo J, Casares de Cal MÁ, et al. Broadening the applicability of a custom multi-platform panel of microhaplotypes: Bio-geographical ancestry inference and expanded reference data. Front Genet. 2020;11:581041.
73. Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, et al. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet. 2014;12:215–24.
74. de la Puente M, Phillips C, Xavier C, Amigo J, Carracedo A, Parson W, et al. Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. Forensic Sci Int Genet. 2020;45:102213.
75. Casanova-Ádan L, Mosquera-Miguel A, González-Bao J, Ambroa-Conde A, Ruiz-Ramírez J, Cabrejas-Olalla A, et al. Adapting an established Ampliseq microhaplotype panel to nanopore sequencing through direct PCR. Forensic Sci Int Genet. 2023;67:102937.
76. Schaffner S. The X chromosome in population genetics. Nat Rev Genet. 2004;5:43–51.
77. Hearne CM, Todd JA. Tetranucleotide repeat polymorphism at the HPRT locus. Nucleic Acids Res. 1991;19(19):5450.
78. Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Leg Med. 2020;45:101709.
79. Bottinelli M, Gouy A, Utz S, Zieger M. Population genetic analysis of 12 X-chromosomal STRs in a Swiss sample. Int J Legal Med. 2022;136(2):561–3.
80. Zhang Y, Yu Z, Mo X, Zhao X, Li W, Liu H, et al. Development and validation of a new 18 X-STR typing assay for forensic applications. Electrophoresis. 2020;42(6):766–73.
81. Ferragut JF, Bentayebi K, Barbaro A, Ramírez M, Saguillo AY, Ramon C, et al. Exploring the western mediterranean through X-chromosome. Int J Legal Med. 2021;135(3):787–90.
82. Garcia FM, Bessa BGO, dos Santos EVW, Pereira JDP, Alves LNR, Vianna LA, et al. Forensic applications of markers present on the X chromosome. Genes (Basel). 2022;13(9):1597.
83. Pinto N, Gusmão L, Amorim A. X-chromosome markers in kinship testing: A generalisation of the IBD approach identifying situations where their contribution is crucial. Forensic Sci Int Genet. 2011;5(1):27–32.
84. Gomes C, Magalhaes M, Alves C, Amorim A, Pinto N, Gusmao L. Comparative evaluation of alternative batteries of genetic markers to complement autosomal STRs in kinship investigations: Autosomal indels vs. X-chromosome STRs. Int J Legal Med. 2012;126(6):917–21.
85. Gomes I, Pinto N, Antão-Sousa S, Gomes V, Gusmão L, Amorim A. Twenty years later: A comprehensive review of the X chromosome use in forensic genetics. Front Genet. 2020;11:926.
86. Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, et al. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science. 1985;230:1403–6.
87. Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet. 1992;89(4):389–94.
88. Roewer L, Epplen JT. Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. Forensic Sci Int. 1992;53(2):163–71.
89. Prinz M, Boll K, Baum H, Shaler B. Multiplexing of Y chromosome specific STRs and performance for mixed samples. Forensic Sci Int. 1997;85(3):209–18.
90. Sinha SK. Forensic casework applications using Y-PLEX™ 6 and Y-PLEX™ 5 systems. Forensic Sci Rev. 2003;15(2):199–203.
91. Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, et al. Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. Forensic Sci Int. 2005;148(1):1–14.
92. Diaz-Lacava A, Walier M, Willuweit S, Wienker TF, Fimmers R, Baur MP, et al. Geostatistical inference of main Y-STR-haplotype groups in Europe. Forensic Sci Int Genet. 2011;5(2):91–4.

93. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, et al. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000;26(3):358–61.
94. Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, et al. Paternal population history of East Asia: Sources, patterns, and microevolutionary processes. Am J Hum Genet. 2001;69(3):615–28.
95. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136(5):621–35.
96. Willuweit S, Roewer L. The new Y chromosome haplotype reference database. Forensic Sci Int Genet. 2015;15:43–8.
97. Greenberg BD, Newbold JE, Sugino A. Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene. 1983;21(1–2):33–49.
98. Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, et al. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 1994;6:130–5.
99. Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA: Applications, debates, and foundations. Annu Rev Genomics Hum Genet. 2003;4:119–41.
100. Tuazon OM. Universal forensic DNA databases : acceptable or illegal under the European Court of Human Rights regime? J Law Biosci. 2021;8(1):lsab022.
101. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015;18:33–48.
102. Albright TD. Why eyewitnesses fail. Proc Natl Acad Sci. 2017;114(30):7758–64.
103. Neufeld P, Scheck B. Innocence Project. Eyewitness misidentification. [Internet]. [cited 2023 Nov 28]. Available from: https://innocenceproject.org/eyewitness-misidentification/
104. The National Registry of Exonerations [Internet]. [cited 2023 Nov 28]. Available from: https://www.law.umich.edu/special/exoneration/Pages/detaillist.aspx?View=%7Bfaf6eddb-5a68-4f8f-8a52-2c61f5bf9ea7%7D&SortField=DNA&SortDir=Asc&FilterField1=MWID&FilterValue1=8_MWID&FilterField2=DNA&FilterValue2=8_DNA
105. Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2011;12:179–92.
106. Kayser M, Branicki W, Parson W, Phillips C. Recent advances in Forensic DNA Phenotyping of appearance, ancestry and age. Forensic Sci Int Genet. 2023;65:102870.
107. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507–15.
108. Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, et al. Genetic structure of human populations. Science. 2002;298(5602):2381–5.
109. Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet. 2015;18:49–65.
110. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat. 2008;29(5):648–58.
111. Times N. Twenty years in prison for 1992 Zaandam murder [Internet]. [cited 2023 Dec 5]. Available from: https://nltimes.nl/2018/12/11/twenty-years-prison-1992-zaandam-murder
112. Kayser M, Schneider PM. DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet. 2009;3:154–61.
113. Frudakis T, Terravainen T, Thomas M. Multilocus OCA2 genotypes specify human iris colors. Hum Genet. 2007;122(3–4):311–26.
114. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007;39(12):1443–52.
115. Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, et al. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014;13:34–44.
116. Walsh S, Liu F, Ballantyne KN, Van Oven M, Lao O, Kayser M. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet. 2011;5(3):170–80.
117. Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, et al. Further development of forensic eye color predictive tests. Forensic Sci Int Genet. 2013;7(1):28–40.
118. Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens ACJW, et al. Eye color and the prediction of complex phenotypes from genotypes. Curr Biol. 2009;19(5):192–3.
119. Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, et al. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet. 2014;9:150–61.
120. Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-kamysz A, et al. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet. 2013;7(1):98–115.
121. Grimes EA, Noake PJ, Dixon L, Urquhart A. Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. Forensic Sci Int. 2001;122:124–9.
122. Branicki W, Liu F, Van Duijn K, Draus-Barini J, Pośpiech E, Walsh S, et al. Model-based prediction of human

hair color using DNA variants. Hum Genet. 2011;129(4):443–54.
123. Söchtig J, Phillips C, Maroñas O, Gómez-Tato A, Cruz R, Alvarez-Dios J, et al. Exploration of SNP variants affecting hair colour prediction in Europeans. Int J Legal Med. 2015;129(5):963–75.
124. Stokowski RP, Pant PVK, Dadd T, Fereday A, Hinds DA, Jarman C, et al. A genomewide association study of skin pigmentation in a South Asian population. Am J Hum Genet. 2007;81:1119–32.
125. Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, et al. Global skin colour prediction from DNA. Hum Genet. 2017;136(7):847–63.
126. Schneider PM, Prainsack B, Kayser M. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. Dtsch Arztebl Int. 2019;116(51–52):873–80.
127. Waddington CH. The Epigenotype. Endeavour. 1942;1:18–20.
128. Mazzio EA, Soliman KFA. Basic concepts of epigenetics impact of environmental signals on gene expression. Epigenetics. 2012;7:119–30.
129. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604–9.
130. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007;128(4):669–81.
131. Allfrey VG, Mirsky AE. Structural modifications of histones and their possible role in the regulation of RNA synthesis. Science. 1964;144(3618):559.
132. Holley RW, Everett GA, Madison JT, Zamir A. Nucleotide sequences in the yeast alanine transfer ribonucleic acid. J Biol Chem. 1965;240:2122–8.
133. Song B, Qian J, Fu J. Research progress and potential application of microRNA and other non-coding RNAs in forensic medicine. Int J Legal Med. 2023;138(2):329–50.
134. Bauer M. RNA in forensic science. Forensic Sci Int Genet. 2007;1(1):69–74.
135. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517(7534):321–6.
136. Hughes AL, Kelley JR, Klose RJ. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. Biochim Biophys Acta Gene Regul Mech. 2020;1863(8):194567.
137. Xu F, Mao C, Ding Y, Rui C, Wu L, Shi A, et al. Molecular and enzymatic profiles of mammalian DNA methyltransferases: structures and targets for drugs. Curr Med Chem. 2010;17(33):4052–71.
138. Neidhart M. DNA methylation and complex human disease. Elsevier Inc.; 2015. 1–8 p.
139. Zhang L, Lu X, Lu J, Liang H, Dai Q, Xu G-L, et al. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA HHS Public Access Author manuscript. Nat Chem Biol. 2012;8(4):328–30.
140. Riggs AD, Jones PA. 5-methylcytosine, gene regulation, and cancer. Adv Cancer Res. 1983;40:1–30.
141. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613(7943):355–64.
142. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Ignacio Martin-Subero J, Rodriguez-Ubreva J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. 2010;20(2):170–9.
143. Maleknia M, Ahmadirad N, Golab F, Katebi Y, Haj Mohamad Ebrahim Ketabforoush A. DNA methylation in cancer: Epigenetic view of dietary and lifestyle factors. Epigenetics Insights. 2023;16.
144. Hikoya, Hayatsu. Yusuke, Wataya, Kazushige K. The addition of sodium bisulfite to uracil and cytosine. J Am Chem Soc. 1970;92(3):724–6.
145. Shapiro R, Servis RE, Welcher M. Reactions of uracil and cytosine derivatives with sodium bisulfite. A specific deamination method. J Am Chem Soc. 1970;92(2):422–4.
146. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci. 1992;89(5):1827–31.
147. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. PLoS One. 2018;13(6):e0199091.
148. Woodcock DM, Crowther PJ, Diver WP. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. Biochem Biophys Res Commun. 1987;145(2):888–94.
149. Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci U S A. 2000;97(10):5237–42.
150. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315–22.
151. Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A. 2009;106(3):671–8.
152. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol. 2009;27(4):353–60.
153. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation

maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766–70.
154. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.
155. Ziller MJ, Stamenova EK, Gu H, Gnirke A, Meissner A. Targeted bisulfite sequencing of the dynamic DNA methylome. Epigenetics Chromatin. 2016;9:55.
156. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1(1):177–200.
157. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95.
158. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8(3):389–99.
159. Kriukienė E, Tomkuvienė M, Klimašauskas S. 5-Hydroxymethylcytosine: the many faces of the sixth base of mammalian DNA. Chem Soc Rev. 2024;53(5):2264–83.
160. Kaur D, Lee SM, Goldberg D, Spix NJ, Hinoue T, Li H-T, et al. Comprehensive evaluation of the Infinium human MethylationEPIC v2 BeadChip. Epigenetics Commun. 2023;3(1):6.
161. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10.000 daltons. Anal Chem. 1988;60:2299–301.
162. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. Protein and polymer analysis up to m/z 100.000 by laser desorption time-of-flight mass spectrometry. Rapid Commun Mass Spectrom. 1988;2:151–3.
163. Tost J, Gut IG. Genotyping single nucleotide polymorphisms by mass spectrometry. Mass Spectrom Rev. 2022;21:388–418.
164. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proc Natl Acad Sci U S A. 2005;102(44):15785–90.
165. Freire-Aradas A, Phillips C, Lareu M. Forensic individual age estimation with DNA: From initial approaches to methylation tests. Forensic Sci Rev. 2017;29(2):121–44.
166. Hyman ED. A new method of sequencing DNA. Anal Biochem. 1988;174(2):423–36.
167. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity. Anal Biochem. 1987;167(2):235–8.
168. Jörg T, Gut IG. DNA methylation analysis by pyrosequencing. Nat Protoc. 2007;2(9):2265–75.
169. Sokolov BP. Primer extension technique for the detection of single nucleotide in genomic DNA. Nucleic Acids Res. 1990;18(12):3671.
170. Makridakis NM, Reichardt JKV. Multiplex automated primer extension analysis: Simultaneous genotyping of template used for primer extension. Biotechniques. 2001;31:1374–80.
171. Uhlmann K, Brinckmann A, Toliat MR, Ritter H, Nürnberg P. Evaluation of a potential epigenetic biomarker by quantitative methyl-single nucleotide polymorphism analysis. Electrophoresis. 2002;23:4072–9.
172. Lee HY, Jung SE, Oh YN, Choi A, Yang WI, Shin KJ. Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study. Forensic Sci Int Genet. 2015;19:28–34.
173. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.
174. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872–6.
175. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14(4):407–10.
176. Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J, Rolf B. Finding the needle in the haystack: Differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet. 2014;9:42–6.
177. Hwa HL, Lin CY, Yu YJ, Linacre A, Lee JCI. DNA identification of monozygotic twins. Forensic Sci Int Genet. 2024;69:102998.
178. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, et al. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PLoS One. 2009;4(8):21–3.
179. Li C, Zhao S, Zhang N, Zhang S, Hou Y. Differences of DNA methylation profiles between monozygotic twins' blood samples. Mol Biol Rep. 2013;40(9):5275–80.
180. Vidaki A, Díez López C, Carnero-Montoro E, Ralf A, Ward K, Spector T, et al. Epigenetic discrimination of identical twins from blood under the forensic scenario. Forensic Sci Int Genet. 2017;31:67–80.
181. Kim JY, Lee HY, Lee SY, Kim SY, Park JL, Lee SD. DNA methylome profiling of blood to identify individuals in a pair of monozygotic twins. Genes and Genomics. 2023;45(10):1273–9.
182. Stewart L, Evans N, Bexon KJ, Van Der Meer DJ, Williams GA. Differentiating between monozygotic twins through DNA methylation-specific high-resolution melt curve analysis. Anal Biochem. 2015;476:36–9.
183. Xu J, Fu G, Yan L, Craig JM, Zhang X, Fu L, et al. LINE-1 DNA methylation: A potential forensic marker for discriminating monozygotic twins. Forensic Sci Int Genet. 2015;19:136–45.

184. Vidaki A, Kalamara V, Carnero-Montoro E, Spector TD, Bell JT, Kayser M. Investigating the epigenetic discrimination of identical twins using buccal swabs, saliva, and cigarette butts in the forensic setting. Genes (Basel). 2018;9(5):252.
185. Morales-Nebreda L, McLafferty FS, Singer BD. DNA methylation as a transcriptional regulator of the immune system. Transl Res. 2019;204:1–18.
186. Allery JP, Telmon N, Mieusset R, Blanc A, Rouge D. Cytological detection of spermatozoa: comparison of three staining methods. J Forensic Sci. 2001;46:349–51.
187. Virkler K, Lednev IK. Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene. Forensic Sci Int. 2009;188(1–3):1–17.
188. Fleming RI, Harbison S. The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. Forensic Sci Int Genet. 2010;4:244–56.
189. Zubakov D, Boersma AWM, Choi Y, Van Kuijk PF, Wiemer EAC, Kayser M. MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. Int J Legal Med. 2010;124(3):217–26.
190. Rocchi A, Chiti E, Maiese A, Turillazzi E, Spinetti I. MicroRNAs: An update of applications in forensic science. Diagnostics. 2020;11(1):32.
191. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.
192. Ng RK, Gurdon JB. Epigenetic inheritance of cell differentiation status. Cell Cycle. 2008;7(9):1173–7.
193. Ohgane J, Yagi S, Shiota K. Epigenetics: The DNA methylation profile of tissue-dependent and differentially methylated regions in cells. Placenta. 2008;29:29–35.
194. Frumkin D, Wasserstrom A, Budowle B, Davidson A. DNA methylation-based forensic tissue identification. Forensic Sci Int Genet. 2011;5(5):517–24.
195. Eipel M, Mayer F, Arent T, Ferreira MRP, Birkhofer C, Gerstenmaier U, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. Aging (Albany NY). 2016;8(5):1034–48.
196. Frobel J, Boži T, Lenz M, Uciechowski P, Han Y, Birkhofer C, et al. Leukocyte counts based on site-specific DNA methylation analysis. Clin Chem. 2017;64(3):566–75.
197. Theda C, Hwang SH, Czajko A, Loke YJ, Leong P, Craig JM. Quantitation of the cellular content of saliva and buccal swab samples. Sci Rep. 2018;8(1):6944.
198. Madi T, Balamurugan K, Bombardi R, Duncan G, Mccord B. The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. Electrophoresis. 2012;33(12):1736–45.
199. Lee HY, Park MJ, Choi A, An JH, Yang WI, Shin KJ. Potential forensic application of DNA methylation profiling to body fluid identification. Int J Legal Med. 2012;126(1):55–62.
200. An JH, Choi A, Shin KJ, Yang WI, Lee HY. DNA methylation-specific multiplex assays for body fluid identification. Int J Legal Med. 2013;127(1):35–43.
201. Antunes J, Silva DSBS, Balamurugan K, Duncan G, Alho CS, McCord B. Forensic discrimination of vaginal epithelia by DNA methylation analysis through pyrosequencing. Electrophoresis. 2016;37(21):2751–8.
202. Balamurugan K, Bombardi R, Duncan G, Mccord B. Identification of spermatozoa by tissue-specific differential DNA methylation using bisulfite modification and pyrosequencing. Electrophoresis. 2014;35:3079–86.
203. Park JL, Kwon OH, Kim JH, Yoo HS, Lee HC, Woo KM, et al. Identification of body fluid-specific DNA methylation markers for use in forensic science. Forensic Sci Int Genet. 2014;13:147–53.
204. Lee HY, An JH, Jung SE, Oh YN, Lee EY, Choi A, et al. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. Forensic Sci Int Genet. 2015;17:17–24.
205. Lee HY, Jung SE, Lee EH, Yang WI, Shin KJ. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. Forensic Sci Int Genet. 2016;24:75–82.
206. Forat S, Huettel B, Reinhardt R, Fimmers R, Haidl G, Denschlag D, et al. Methylation markers for the identification of body fluids and tissues from forensic trace evidence. PLoS One. 2016;11(2):e0147973.
207. Alghanim H, Balamurugan K, Mccord B. Development of DNA methylation markers for sperm, saliva and blood indentification using pyrosequencing and qPCR/HRM. Anal Biochem. 2020;611:113933.
208. Rothe J, Becker JM, Charchinezhadamouei M, Mähr S, Lembeck F, Dannemann N, et al. Expanding the scope of methylation-sensitive restriction enzyme (MSRE) PCR for forensic identification of body fluids through the novel use of methylation-dependent restriction enzymes (MDRE) and the combination of autosomal and Y-chromosomal markers. Int J Legal Med. 2024;138(2):375–93.
209. Wasserstrom A, Frumkin D, Davidson A, Shpitzen M, Herman Y, Gafny R. Demonstration of DSI-semen - A novel DNA methylation-based forensic semen identification assay. Forensic Sci Int Genet. 2013;7(1):136–42.
210. Vidaki A, Giangasparo F, Syndercombe Court D. Discovery of potential DNA methylation markers for forensic tissue identification using bisulphite pyrosequencing. Electrophoresis. 2016;37(21):2767–79.
211. Lee J, Lee J, Naue J, Fleckhaus J, Freire-Aradas A, Neubauer J, et al. A collaborative exercise on DNA methylation-based age prediction and body fluid typing. Forensic Sci Int Genet. 2022;57:102656.

212. Kim BM, Park SU, Schmelzer L, Yang S-B, Lee SD, Kim M-Y, et al. DNA methylation-based organ tissue identification: Marker identification, SNaPshot multiplex assay development, and interlaboratory comparison. Forensic Sci Int Genet. 2024;71:103052.
213. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153(6):1194–217.
214. Cevenini E, Invidia L, Lescai F, Salvioli S, Tieri P, Castellani G, et al. Human models of aging and longevity. Expert Opin Biol Ther. 2008;8(9):1393–405.
215. Candore G, Balistreri CR, Listì F, Grimaldi MP, Vasto S, Colonna-Romano G, et al. Immunogenetics, gender, and longevity. Ann N Y Acad Sci. 2006;1089(1):516–37.
216. Moskalev AA, Shaposhnikov M V, Plyusnina EN, Zhavoronkov A, Budovsky A, Yanai H, et al. The role of DNA damage and repair in aging through the prism of Koch-like criteria. Ageing Res Rev. 2013;12(2):661–84.
217. Park CB, Larsson N-G. Mitochondrial DNA mutations in disease and aging. J Cell Biol. 2011;193(5):809–18.
218. Linnane AW, Marzuki S, Ozawa T, Tanaka M. Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases. Lancet (London, England). 1989;1(8639):642–5.
219. Green DR, Galluzzi L, Kroemer G. Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. Science. 2011;333(6046):1109–12.
220. Fontana L, Partridge L, Longo VD. Dietary restriction, growth factors and aging: from yeast to humans. Science. 2010;328(5976):321–6.
221. Blackburn EH, Greider CW, Szostak JW. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nat Med. 2006;12(10):1133–8.
222. Talens RP, Christensen K, Putter H, Willemsen G, Christiansen L, Kremer D, et al. Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. Aging Cell. 2012;11(4):694–703.
223. Freije JMP, López-Otín C. Reprogramming aging and progeria. Curr Opin Cell Biol. 2012;24(6):757–64.
224. Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE. Biological and chemical approaches to diseases of proteostasis deficiency. Annu Rev Biochem. 2009;78:959–91.
225. Rando TA, Chang HY. Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock. Cell. 2012;148(1–2):46–57.
226. Salminen A, Kaarniranta K, Kauppinen A. Inflammaging: disturbed interplay between autophagy and inflammasomes. Aging (Albany NY). 2012;4(3):166–75.
227. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. Cell. 2023;186(2):243–78.
228. Scheuer L. Application of osteology to forensic medicine. Clin Anat. 2002;15(4):297–312.
229. Wittschieber D, Ottow C, Vieth V, Küppers M, Schulz R, Hassu J, et al. Projection radiography of the clavicle: still recommendable for forensic age diagnostics in living individuals? Int J Legal Med. 2015;129(1):187–93.
230. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke KD, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med. 2004;118(1):5–8.
231. Olze A, Van Niekerk P, Ishikawa T, Zhu BL, Schulz R, Maeda H, et al. Comparative study on the effect of ethnicity on wisdom tooth eruption. Int J Legal Med. 2007;121(6):445–8.
232. Michikawa Y, Mazzucchelli F, Bresolin N, Scarlato G, Attardi G. Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. Science. 1999;286(5440):774–9.
233. Meissner C, Ritz-Timme S. Molecular pathology and age estimation. Forensic Sci Int. 2010;203(1–3):34–43.
234. Shay JW. Telomeres and aging. Curr Opin Cell Biol. 2018;52:1–7.
235. Pilin A, Pudil F, Bencko V. Changes in colour of different human tissues as a marker of age. Int J Legal Med. 2007;121(2):158–62.
236. Ohtani S, Yamamoto T. Age estimation by amino acid racemization in human teeth. J Forensic Sci. 2010;55(6):1630–3.
237. Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM, et al. Estimating human age from T-cell DNA rearrangements. Curr Biol. 2010;20(22):R970–1.
238. Wilson VL, Jones PA. DNA methylation decreases in aging but not in immortal cells. Science. 1983;220(4601):1055–7.
239. Cooney CA. Are somatic cells inherently deficient in methylation metabolism? A proposed mechanism for DNA methylation loss, senescence and aging. Growth Dev Aging. 1993;57(4):261–73.
240. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. Aging Cell. 2015;14(6):924–32.
241. Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, et al. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 2010;20(3):332–40.
242. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20(4):434–9.
243. Marttila S, Kananen L, Häyrynen S, Jylhävä J, Nevalainen T, Hervonen A, et al. Ageing-associated changes in

the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics. 2015;16:179.
244. Martin GM. Epigenetic drift in aging identical twins. Proc Natl Acad Sci. 2005;102(30):10413–4.
245. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
246. Feng L, Peng F, Li S, Jiang L, Sun H, Ji A, et al. Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. Forensic Sci Int Genet. 2018;35:38–46.
247. Hong S, Shin K, Jung S, Lee E, Lee H. Platform-independent models for age prediction using DNA methylation data. Forensic Sci Int Genet. 2019;38:39–47.
248. So MH, Lee HY. Genetic analyzer-dependent DNA methylation detection and its application to existing age prediction models. Electrophoresis. 2021;42(14–15):1497–506.
249. Freire-Aradas A, Pośpiech E, Aliferi A, Girón-Santamaría L, Mosquera-Miguel A, Pisarek A, et al. A comparison of forensic age prediction models using data from four DNA methylation technologies. Front Genet. 2020;11:932.
250. So MH, Lee JE, Lee HY. Strategies to deal with genetic analyzer-specific DNA methylation measurements. Electrophoresis. 2024;45:906–15.
251. Vittinghoff E, Glidden D V., Shiboski SC, McCulloch CE. Regression methods in biostatistics. 2nd ed. Springer New York, NY; 2012. 509 p.
252. Kresovich JK, Lopez AMM, Garval EL, Xu Z, White AJ, Dale P, et al. Alcohol consumption and methylation-based measures of biological age. J Gerontol Ser A Biol Sci Med Sci. 2021;76(12):2107–11.
253. Galkin F, Kovalchuk O, Koldasbayeva D, Zhavoronkov A, Bischof E. Stress, diet, exercise: Common environmental factors and their impact on epigenetic age. Ageing Res Rev. 2023;88:101956.
254. Aurich S, Müller L, Kovacs P, Keller M. Implication of DNA methylation during lifestyle mediated weight loss. Front Endocrinol (Lausanne). 2023;14:1181002.
255. Gao X, Zhang Y, Breitling LP, Brenner H. Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget. 2016;7(30):46878–89.
256. Stephenson M, Bollepalli S, Cazaly E, Salvatore JE, Street WF. Associations of alcohol consumption with epigenome-wide DNA methylation and epigenetic age acceleration: individual-level and co-twin comparison analyses. Alcohol Clin Exp Res. 2022;45(2):318–28.
257. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10(4):573–91.
258. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11(2):303–27.
259. McGreevy KM, Radak Z, Torma F, Jokai M, Lu AT, Belsky DW, et al. DNAmFitAge: biological age indicator incorporating physical fitness. Aging (Albany NY). 2023;15(10):3904–38.
260. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. Genome Biol. 2019;20(1):249.
261. Jenkins TG, Aston KI, Cairns B, Smith A, Carrell DT. Paternal germ line aging: DNA methylation age prediction from human sperm. BMC Genomics. 2018;19(1):763.
262. Aliferi A, Sundaram S, Ballard D, Freire-Aradas A, Phillips C, Lareu MV, et al. Combining current knowledge on DNA methylation-based age estimation towards the development of a superior forensic DNA intelligence tool. Forensic Sci Int Genet. 2022;57:102637.
263. Horvath S. DNA methylation age of human tissues and cell types. Genome biol. 2013;14(10):R115.
264. Heidegger A, Pisarek A, de la Puente M, Niederstätter H, Pośpiech E, Woźniak A, et al. Development and inter-laboratory validation of the VISAGE enhanced tool for age estimation from semen using quantitative DNA methylation analysis. Forensic Sci Int Genet. 2022;56:102596.
265. 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):953–61.
266. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.
267. Weidner CI, 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.
268. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Z, 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.
269. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Z, 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.
270. Park J-L, Kim JH, Seo E, Bae DH, Kim S-Y, Lee H-C, et al. Identification and evaluation of age-correlated DNA methylation markers for forensic use. Forensic Sci Int Genet. 2016;23:64–70.
271. 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.
272. 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:e249–51.
273. 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;28:225–36.
274. Naue J, Hoefsloot HCJ, Mook ORF, Rijlaarsdam-hoekstra L, Zwalm MCH Van Der, 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.
275. Jung S-E, Min S, Rom S, Hee E, Shin K, Young H. 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.
276. Wózniak A, Heidegger A, Piniewska-Róg D, Pospiech E, Xavier C, Pisarek A, et al. Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. Aging (Albany NY). 2021;13(5):6459–84.
277. Han X, Xiao C, Yi S, Li Y, Chen M, Huang D. Accurate age estimation from blood samples of Han Chinese individuals using eight high-performance age-related CpG sites. Int J Legal Med. 2022;136(6):1655–65.
278. Chen D, Chao DL, Rocha L, Kolar M, Nguyen Huu VA, Krawczyk M, et al. The lipid elongation enzyme ELOVL2 is a molecular regulator of aging in the retina. Aging Cell. 2020;19(2):e13100.
279. Chao DL, Skowronska-Krawczyk D. ELOVL2: Not just a biomarker of aging. Transl Med Aging. 2020;4:78–80.
280. 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;17:129–36.
281. Freire-Aradas A, Phillips C, Girón-santamaría L, Mosquera-miguel A, Gómez-tato A, Casares MÁ, et al. Tracking age-correlated DNA methylation markers in the young. Forensic Sci Int Genet. 2018;36:50–9.
282. Bekaert B, Kamalandua A, Zapico SC, Voorde W Van De, Bekaert B. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. Epigenetics. 2015;10(10):922–30.
283. 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):465–70.
284. Correia Dias H, Cunha E, Corte Real F, Manco L. Age prediction in living: Forensic epigenetic age estimation based on blood samples. Leg Med. 2020;47:101763.
285. 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.
286. Vidaki A, González DM, Jiménez BP, Kayser M. Male-specific age estimation based on Y-chromosomal DNA methylation. Aging (Albany NY). 2021;13(5):6442–58.
287. Jiang L, Zhang K, Wei X, Li J, Wang S, Wang Z, et al. Developing a male-specific age predictive model based on Y-CpGs for forensic analysis. Forensic Sci Int. 2023;343:111566.
288. Thiede C, Prange-Krex G, Freiberg-Richter J, Bornhäuser 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 Transpl. 2000;25(5):575–7.
289. Bocklandt S, Lin W, Sehl ME, Sa FJ, Sinsheimer JS, Horvath S, et al. Epigenetic predictor of age. PLoS One. 2011;6(6):e14821.
290. Hong SR, Jung S, Lee EH, Shin K, 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.
291. Schwender K, Holländer O, Klopfleisch S, Eveslage M, Danzer MF, Pfeiffer H, et al. Development of two age estimation models for buccal swab samples based on 3 CpG sites analyzed with pyrosequencing and minisequencing. Forensic Sci Int Genet. 2021;53:102521.
292. Ho Lee M, Hee Hwang J, Min Seong K, Jin Ahn J, Jun Kim S, Yong Hwang S, et al. Application of droplet digital PCR method for DNA methylation-based age prediction from saliva. Leg Med. 2022;54:101992.
293. Marcante B, Delicati A, Onofri M, Tozzo P, Caenazzo L. Estimation of human chronological age from buccal swab samples through a DNA methylation analysis approach of a five-locus multiple regression model. Int J Mol Sci. 2024;25(2):935.
294. 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(5):10444.
295. 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.
296. Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT. Age-associated sperm DNA methylation alterations:

Possible implications in offspring disease susceptibility. PLOS Genet. 2014;10(7):e1004458.
297. Jenkins TG, James ER, Aston KI, Salas-Huetos A, Pastuszak AW, Smith KR, et al. Age-associated sperm DNA methylation patterns do not directly persist trans-generationally. Epigenetics Chromatin. 2019;12(1):74.
298. Kotková L, Drábek J. Age-related changes in sperm DNA methylation and their forensic and clinical implications. Epigenomics. 2023;15(21):1157–73.
299. Long S, Kenworthy S. Round cells in diagnostic semen analysis: A guide for laboratories and clinicians. Br J Biomed Sci. 2022;79:10129.
300. Li L, Song F, Huang Y, Zhu H, Hou Y. Age-associated DNA methylation determination of semen by pyrosequencing in Chinese Han population. Forensic Sci Int Genet Suppl Ser. 2017;6:e99–100.
301. Pisarek A, Pospiech E, Heidegger A, Xavier C, Papiez A, Piniewska-Róg D, et al. Epigenetic age prediction in semen - marker selection and model development. Aging (Albany NY). 2021;13(15):19145–64.
302. Li L, Song F, Lang M, Hou J, Wang Z, Prinz M, et al. Methylation-based age prediction using pyrosequencing platform from seminal stains in Han chinese males. J Forensic Sci. 2019;65(2):610–9.
303. Lee JE, Park SU, So MH, Lee HY. Age prediction using DNA methylation of Y-chromosomal CpGs in semen samples. Forensic Sci Int Genet. 2024;69:103007.
304. Márquez-Ruiz AB, González-Herrera L, Luna J de D, Valenzuela A. DNA methylation levels and telomere length in human teeth: usefulness for age estimation. Int J Legal Med. 2020;134(2):451–9.
305. Hao T, Guo J, Liu J, Wang J, Liu Z, Cheng X, et al. Predicting human age by detecting DNA methylation status in hair. Electrophoresis. 2021;42(11):1255–61.
306. C Zapico S, Gauthier Q, Antevska A, McCord BR. Identifying methylation patterns in dental pulp aging: Application to age-at-death estimation in forensic anthropology. Int J Mol Sci. 2021;22(7):3717.
307. Fokias K, Dierckx L, Van de Voorde W, Bekaert B. Improving the age estimation model for toenails. Forensic Sci Int Genet. 2023;66:102911.
308. Naue J, Sänger 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;36:152–9.
309. Lee HY, Hong SR, Lee JE, Hwang IK, Kim NY, Lee JM, et al. Epigenetic age signatures in bones. Forensic Sci Int Genet. 2020;46:102261.
310. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY). 2018;10(7):1758–75.
311. Correia Dias H, Manco L, Corte Real F, Cunha E. A blood-bone-tooth model for age prediction in forensic contexts. Vol. 10, Biology. 2021. p. 1312.
312. 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.
313. Hillson S. Teeth. 2nd ed. Cambridge Manuals in Archaeology. Cambridge: Cambridge University Press; 2005.
314. Naue J, Winkelmann J, Schmidt U, Lutz-Bonengel S. Analysis of age-dependent DNA methylation changes in plucked hair samples using massive parallel sequencing. Rechtsmedizin. 2021;31(3):226–33.
315. Li C, Li Y, Zhou G, Gao Y, Ma S, Chen Y, et al. Whole-genome bisulfite sequencing of goat skins identifies signatures associated with hair cycling. BMC Genomics. 2018;19(1):638.
316. Fokias K, Dierckx L, Van de Voorde W, Bekaert B. Age determination through DNA methylation patterns in fingernails and toenails. Forensic Sci Int Genet. 2023;64:102846.
317. Voisin S, Jacques M, Landen S, Harvey NR, Haupt LM, Griffiths LR, et al. Meta-analysis of genome-wide DNA methylation and integrative omics of age in human skeletal muscle. J Cachexia Sarcopenia Muscle. 2021;12(4):1064–78.
318. Lee JM, Park SU, Lee SD, Lee HY. Application of array-based age prediction models to post-mortem tissue samples. Forensic Sci Int Genet. 2024;68:102940.
319. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018;14(8):e1007544.
320. INE. Infografía día mundial sin tabaco [Internet]. Instituto Nacional de Estadística (INE). 2023 [cited 2024 Mar 18]. Available from: https://www.ine.es/infografias/infografia_tabaco.pdf
321. OEDA. Informe 2023. Alcohol, tabaco y drogas ilegales en España. [Internet]. Observatorio español de las drogas y las adicciones. 2023 [cited 2024 Mar 18]. Available from: https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/informesEstadisticas/pdf/2023OEDA-INFORME.pdf
322. Eurostat. Eurostat. Tobacco consumption statistics [Internet]. European health interview survey. 2019 [cited 2024 Mar 18]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Tobacco_consumption_statistics#Level_of_cigarette_consumption
323. Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132.
324. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA

methylation: 27K discovery and replication. Am J Hum Genet. 2011;88(4):450–7.
325. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8(5):e63812.
326. Gao X, Zhang Y, Breitling LP, Brenner H. Tobacco smoking and methylation of genes related to lung cancer development. Oncotarget. 2016;7(37):59017–28.
327. OMS. Las 10 principales causas de defunción [Internet]. Organización Mundial de la Salud. 2020 [cited 2024 Mar 18]. Available from: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death
328. Dogan M, Lei M, Beach S, Brody G, Philbert R. Alcohol and tobacco consumption alter hypothalamic pituitary adrenal axis DNA methylation. Psychoneuroendocrinology. 2016;66:176–84.
329. Gao X, Zhang Y, Saum KU, Schöttker B, Breitling LP, Brenner H. Tobacco smoking and smoking-related DNA methylation are associated with the development of frailty among older adults. Epigenetics. 2017;12(2):149–56.
330. Cosin-Tomas M, Cilleros-Portet A, Aguilar-Lacasaña S, Fernandez-Jimenez N, Bustamante M. Prenatal maternal smoke, DNA methylation, and multi-omics of tissues and child health. Curr Environ Heal Reports. 2022;9(3):502–12.
331. Carreras-Gallo N, Dwaraka VB, Cáceres A, Smith R, Mendez TL, Went H, et al. Impact of tobacco, alcohol, and marijuana on genome-wide DNA methylation and its relationship with hypertension. Epigenetics. 2023;18(1):2214392.
332. Si J, Chen L, Yu C, Guo Y, Sun D, Pang Y, et al. Healthy lifestyle, DNA methylation age acceleration, and incident risk of coronary heart disease. Clin Epigenetics. 2023;15:52.
333. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao Y, Jones A, et al. Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. JAMA Oncol. 2015;1(4):476–85.
334. Zhang Y, Schöttker B, Florath I, Stock C, Butterbach K, Holleczek B, et al. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environ Health Perspect. 2016;124(1):67–74.
335. Eckhardt CM, Wu H, Prada D, Vokonas PS, Sparrow D, Hou L, et al. Predicting risk of lung function impairment and all-cause mortality using a DNA methylation-based classifier of tobacco smoke exposure. Respir Med. 2022;200:106896.
336. Shenker NS, Ueland PM, Polidoro S, Van Veldhoven K, Ricceri F, Brown R, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology. 2013;24(5):712–6.
337. Philibert R. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol. 2015;6:656.
338. Zhang Y, Florath I, Saum K, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395–403.
339. Kondratyev N, Golov A, Alfimova M, Lezheiko T, Golimbet V. Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation. Clin Epigenetics. 2018;10(1):130.
340. Chamberlain JD, Nusslé S, Chapatte L, Kinnaer C, Petrovic D, Pradervand S, et al. Blood DNA methylation signatures of lifestyle exposures: tobacco and alcohol consumption. Clin Epigenetics. 2022;14(1):155.
341. Tsaprouni L, Yang T, Bell J, Dick K, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382–96.
342. Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet. 2015;24(8):2349–59.
343. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8(5):599–618.
344. Dugué P, Jung C, Joo JE, Wang X, Ming E, Makalic E, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. Epigenetics. 2020;15(4):358–68.
345. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
346. Alghanim H, Wu W, Mccord B. DNA methylation assay based on pyrosequencing for determination of smoking status. Electrophoresis. 2018;39(21):2806–14.
347. Maas SCE, Vidaki A, Wilson R, Teumer A, Liu F, Meurs JBJ Van. Validated inference of smoking habits from blood with a finite DNA methylation marker set. Eur J Epidemiol. 2019;34(11):1055–74.
348. Philibert R, Mills JA, Dogan M, Beach SRH, Long JD. AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. Am J Med Genet Part B Neuropsychiatr Genet. 2020;183(1):51–60.
349. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
350. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrel MA, et al. Epigenome-wide

association study in European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843–51.
351. McCartney DL, Stevenson AJ, Hillary RF, Walker RM, Bermingham ML, Morris SW, et al. Epigenetic signatures of starting and stopping smoking. EBioMedicine. 2018;37:214–20.
352. Thomasson HR, Edenberg HJ, Crabb DW, Mai X, Jerome RE, Li T, et al. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in chinese men. Am J Hum Genet. 1991;48(4):677–81.
353. Chen C, Lu R, Chen Y, Wang M, Chang Y, Li T, et al. Interaction between the functional polymorphisms of the alcohol- metabolism genes in protection against alcoholism. Am J Hum Genet. 1999;65(3):795–807.
354. Tyndale RF. Genetics of alcohol and tobacco use in humans. Ann Med. 2003;35(8):94–121.
355. Liu I, Blacker D, Xu R, Fitzmaurice G, Lyons M, Tsuang M. Genetic and environmental contributions to the development of alcohol dependence in male twins. Arch Gen Psychiatry. 2004;61(9):897–903.
356. Hoenicka J, Ampuero I, Ramos Atance JA. Aspectos genéticos del alcoholismo. Trastor Adict. 2003;5(3):213–22.
357. Corrao G, Ph D, Bagnardi V, Sc D, Zambon A, Sc D, et al. A meta-analysis of alcohol consumption and the risk of 15 diseases. Prev Med (Baltim). 2004;38(5):613–9.
358. Lee HY, Lee SD, Shin KJ. Forensic DNA methylation profiling from evidence material for investigative leads. BMB Rep. 2016;49(7):359–69.
359. Bönsch D, Lenz B, Reulbach U, Kornhuber J, Bleich S. Homocysteine associated genomic DNA hypermethylation in patients with chronic alcoholism. J Neural Transm. 2004;111(12):1611–6.
360. Philibert RA, Plume JM, Gibbons FX, Brody GH, Beach SRH. The impact of recent alcohol use on genome wide DNA methylation signatures. Front Genet. 2012;3:54.
361. Weng JT, Wu LS, Lee C, Hsu PW, Cheng ATA. Integrative epigenetic profiling analysis identifies DNA methylation changes associated with chronic alcohol consumption. Comput Biol Med. 2015;64:299–306.
362. Zhang R, Miao Q, Wang C, Zhao R, Li W, Haile CN, et al. Genome-wide DNA methylation analysis in alcohol dependence. Addict Biol. 2013;18(2):392–403.
363. Wilson LE, Xu Z, Harlid S, White AJ, Troester MA, Sandler DP, et al. Alcohol and DNA methylation: An epigenome-wide association study in blood and normal breast tissue. Am J Epidemiol. 2019;188(6):1055–65.
364. Dugué PA, Wang X, Baglietto L, Wilson R, Lehne B, Makalic E, et al. Alcohol consumption is associated with widespread changes in blood DNA methylation: Analysis of cross-sectional and longitudinal data. Addict Biol. 2019;26(1):e12855.
365. Zhao R, Zhang R, Li W, Liao Y, Tang J, Miao Q, et al. Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence. Asia -Pacific Psychiatry. 2013;5(1):39–50.
366. Philibert RA, Penaluna B, White T, Shires S, Gunter T, Liesveld J, et al. A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs. Epigenetics. 2014;9(9):1212–9.
367. Liangpunsakul S, Lai X, Ross RA, Yu Z, Modlik E, Westerhold C, et al. Novel serum biomarkers for detection of excessive alcohol use. Alcohol Clin Exp Res. 2015;39(3):556–65.
368. INE. Determinantes de salud (consumo de tabaco, exposición pasiva al humo de tabaco, alcohol, problemas medioambientales en la vivienda) [Internet]. Instituto Nacional de Estadística (INE). 2023 [cited 2024 Mar 22]. Available from: https://www.ine.es/ss/Satellite?c=INESeccion_C&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&cid=1259926698156&L=1
369. Maas SCE, Vidaki A, Teumer A, Costeira R, Wilson R, van Dongen J, et al. Validating biomarkers and models for epigenetic inference of alcohol consumption from blood. Clin Epigenetics. 2021;13(1):198.
370. Frezza M, di Padova C, Pozzato G, Terpin M, Baraona E, Lieber C. High blood alcohol levels in women. The role of decreased gastric alcohol dehydrogenase activity and first-pass metabolism. N Engl J Med. 1990;322(2):95–9.
371. Marmot M. Inequality, deprivation and alcohol use. Addiction. 1997;92:13–20.
372. Wong CCY, Mill J, Fernandes C. Drugs and addiction: An introduction to epigenetics. Addiction. 2009;106(3):480–9.
373. Stewart AF, Fulton SL, Maze I. Epigenetics of drug addiction. Cold Spring Harb Perspect Med. 2021;11(7):a040253.
374. Feng J, Shao N, Szulwach KE, Vialou V, Huynh J, Zhong C, et al. Role of Tet1 and 5-hydroxymethylcytosine in cocaine action. Nat Neurosci. 2015;18(4):536–44.
375. Levenson JM, Sweatt JD. Epigenetic mechanisms in memory formation. Nat Rev Neurosci. 2005;6(2):108–18.
376. Smith A, Kaufman F, Sandy MS, Cardenas A. Cannabis exposure during critical windows of development: Epigenetic and molecular pathways implicated in neuropsychiatric disease. Curr Environ Heal reports. 2020;7(3):325–42.
377. Liu J, Chen J, Ehrlich S, Walton E, White T, Perrone-Bizzozero N, et al. Methylation patterns in whole blood correlate with symptoms in schizophrenia patients. Schizophr Bull. 2014;40(4):769–76.
378. Wiedmann M, Kuitunen-Paul S, Basedow LA, Wolff M, DiDonato N, Franzen J, et al. DNA methylation

changes associated with cannabis use and verbal learning performance in adolescents: an exploratory whole genome methylation study. Transl Psychiatry. 2022;12(1):317.
379. Domingos LB, Silva NR, Chaves Filho AJM, Sales AJ, Starnawska A, Joca S. Regulation of DNA methylation by cannabidiol and its implications for psychiatry: New insights from in vivo and in silico models. Genes (Basel). 2022;13(11):2165.
380. Rotter A, Bayerlein K, Hansbauer M, Weiland J, Sperling W, Kornhuber J, et al. CB1 and CB2 receptor expression and promoter methylation in patients with cannabis dependence. Eur Addict Res. 2013;19(1):13–20.
381. Osborne AJ, Pearson JF, Noble AJ, Gemmell NJ, Horwood LJ, Boden JM, et al. Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort. Transl Psychiatry. 2020;10(1):114.
382. Markunas CA, Hancock DB, Xu Z, Quach BC, Fang F, Sandler DP, et al. Epigenome-wide analysis uncovers a blood-based DNA methylation biomarker of lifetime cannabis use. Am J Med Genet B Neuropsychiatr Genet. 2021;186(3):173–82.
383. Nannini DR, Zheng Y, Joyce BT, Kim K, Gao T, Wang J, et al. Genome-wide DNA methylation association study of recent and cumulative marijuana use in middle aged adults. Mol Psychiatry. 2023;28(6):2572–82.
384. Garrett ME, Dennis MF, Bourassa KJ, Hauser MA, Kimbrel NA, Beckham JC, et al. Genome-wide DNA methylation analysis of cannabis use disorder in a veteran cohort enriched for posttraumatic stress disorder. Psychiatry Res. 2024;333:115757.
385. Murphy SK, Itchon-Ramos N, Visco Z, Huang Z, Grenier C, Schrott R, et al. Cannabinoid exposure and altered DNA methylation in rat and human sperm. Epigenetics. 2018;13(12):1208–21.
386. Andersen A, Gerrard M, Gibbons FX, Beach SRH, Philibert R. An examination of risk factors for tobacco and cannabis smoke exposure in adolescents using an epigenetic biomarker. Front psychiatry. 2021;12:688384.
387. Nielsen DA, Yuferov V, Hamon S, Jackson C, Ho A, Ott J, et al. Increased OPRM1 DNA methylation in lymphocytes of methadone-maintained former heroin addicts. Neuropsychopharmacology. 2009;34(4):867–73.
388. Nielsen DA, Utrankar A, Reyes JA, Simons DD, Kosten TR. Epigenetics of drug abuse: predisposition or response. Pharmacogenomics. 2012;13(10):1149–60.
389. Ebrahimi G, Asadikaram G, Akbari H, Nematollahi MH, Abolhassani M, Shahabinejad G, et al. Elevated levels of DNA methylation at the OPRM1 promoter region in men with opioid use disorder. Am J Drug Alcohol Abuse. 2018;44(2):193–9.
390. Camilo C, Maschietto M, Vieira HC, Tahira AC, Gouveia GR, Feio Dos Santos AC, et al. Genome-wide DNA methylation profile in the peripheral blood of cocaine and crack dependents. Rev Bras Psiquiatr. 2019;41(6):485–93.
391. Montalvo-Ortiz JL, Cheng Z, Kranzler HR, Zhang H, Gelernter J. Genomewide study of epigenetic biomarkers of opioid dependence in european-american women. Sci Rep. 2019;9(1):4660.
392. Shu C, Jaffe AE, Sabunciyan S, Ji H, Astemborski J, Sun J, et al. Epigenome-wide association analyses of active injection drug use. Drug Alcohol Depend. 2022;235:109431.
393. Nielsen DA, Hamon S, Yuferov V, Jackson C, Ho A, Ott J, et al. Ethnic diversity of DNA methylation in the OPRM1 promoter region in lymphocytes of heroin addicts. Hum Genet. 2010;127(6):639–49.
394. Khan KM, Thompson AM, Blair SN, Sallis JF, Powell KE, Bull FC, et al. Sport and exercise as contributors to the health of nations. Lancet. 2012;380(9836):59–64.
395. Dimauro I, Sgura A, Pittaluga M, Magi F, Fantini C, Mancinelli R, et al. Regular exercise participation improves genomic stability in diabetic patients: an exploratory study to analyse telomere length and DNA damage. Sci Rep. 2017;7(1):4137.
396. Grazioli E, Cerulli C, Dimauro I, Moretti E, Murri A, Parisi A. New strategy of home-based exercise during pandemic COVID-19 in breast cancer patients: A case study. Vol. 12, Sustainability. 2020. p. 6940.
397. Blair SN, Kohl III HW, Paffenbarger Jr RS, Clark DG, Cooper KH, Gibbons LW. Physical fitness and all-cause mortality: A prospective study of healthy men and women. JAMA. 1989;262(17):2395–401.
398. Barrès R, Yan J, Egan B, Treebak JT, Rasmussen M, Fritz T, et al. Acute exercise remodels promoter methylation in human skeletal muscle. Cell Metab. 2012;15(3):405–11.
399. Nitert MD, Dayeh T, Volkov P, Elgzyri T, Hall E, Nilsson E, et al. Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. Diabetes. 2012;61(12):3322–32.
400. Grazioli E, Dimauro I, Mercatelli N, Wang G, Pitsiladis Y, Di Luigi L, et al. Physical activity in the prevention of human diseases: role of epigenetic modifications. BMC Genomics. 2017;18(8):802.
401. Sellami M, Bragazzi N, Prince MS, Denham J, Elrayess M. Regular, intense exercise training as a healthy aging lifestyle strategy: Preventing DNA damage, telomere shortening and adverse DNA methylation changes over a lifetime. Front Genet. 2021;12:652497.
402. da Silva Rodrigues G, Noronha NY, Noma IHY, de Lima JGR, da Silva Sobrinho AC, de Souza Pinhel MA, et al. 14-Week exercise training modifies the DNA methylation levels at gene sites in non-Alzheimer's disease women aged 50 to 70 years. Exp Gerontol. 2024;186:112362.
403. Moulton C, Murri A, Benotti G, Fantini C, Duranti G, Ceci R, et al. The impact of physical activity on promoter-

specific methylation of genes involved in the redox-status and disease progression: A longitudinal study on post-surgery female breast cancer patients undergoing medical treatment. Redox Biol. 2024;70:103033.
404. de Toro-Martín J, Guénard F, Tchernof A, Hould F-S, Lebel S, Julien F, et al. Body mass index is associated with epigenetic age acceleration in the visceral adipose tissue of subjects with severe obesity. Clin Epigenetics. 2019;11(1):172.
405. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schönfels W, Ahrens M, et al. Obesity accelerates epigenetic aging of human liver. Proc Natl Acad Sci. 2014;111(43):15538–43.
406. Lundgren S, Kuitunen S, Pietiläinen KH, Hurme M, Kähönen M, Männistö S, et al. BMI is positively associated with accelerated epigenetic aging in twin pairs discordant for BMI. J Intern Med. 2021;292(4):627–40.
407. Voisin S, Seale K, Jacques M, Landen S, Harvey NR, Haupt LM, et al. Exercise is associated with younger methylome and transcriptome profiles in human skeletal muscle. Aging Cell. 2024;23(1):e13859.
408. Caporossi D, Dimauro I. Exercise-induced redox modulation as a mediator of DNA methylation in health maintenance and disease prevention. Free Radic Biol Med. 2024;213:113–22.
409. Choi S-W, Friso S. Epigenetics: A new bridge between nutrition and health. Adv Nutr. 2010;1(1):8–16.
410. Hore TA. Modulating epigenetic memory through vitamins and TET: implications for regenerative medicine and cancer treatment. Epigenomics. 2017;9(6):863–71.
411. Arora I, Sharma M, Tollefsbol TO. Combinatorial epigenetics impact of polyphenols and phytochemicals in cancer prevention and therapy. Int J Mol Sci. 2019;20(18):4567.
412. Lim U, Song M-A. Dietary and lifestyle factors of DNA methylation. Methods Mol Biol. 2012;863:359–76.
413. Pauwels S, Ghosh M, Duca RC, Bekaert B, Freson K, Huybrechts I, et al. Maternal intake of methyl-group donors affects DNA methylation of metabolic genes in infants. Clin Epigenetics. 2017;9:16.
414. Guía para el uso forense del ADN. [Madrid]: Comisión Nacional para el uso forense del ADN. Ministerio de Justica; 2019.
415. Bjornsson HT, Fallin MD, Feinberg AP. An integrated epigenetic and genetic approach to common human disease. Trends Genet. 2004;20(8):350–8.
416. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, Kong A, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet. 2011;43(6):561–4.
417. Silva DSBS, Ecker H, Walcott J, Weeden R, Medina A, Gorson JM. Analysis of DNA methylation markers for tissue identification in individuals with different clinical phenotypes. Electrophoresis. 2023;44(13–14):1037–46.
418. Team R Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2020.

ANEXO I

ANEXO I

Este trabajo no ha sido publicado.

Anexo I: DNA methylation data transformation for the application of a cross-platform age prediction model

Abstract:

In a previous study, an age prediction model was developed for saliva and buccal cell samples using DNA methylation data generated with an ABI3130 genetic analyzer. Based on these data, a quantile regression model was built using 7 age-correlated CpGs with a median absolute error (MAE) of ± 3.66 years, placing the corresponding model in the open-access *Snipper* forensic classifier. To adjust for differences in signal balance, which altered the peak height ratios between dyes for each genetic analyzer type, it was necessary to apply data transformation to the updated data generated with the latest platforms when analyzed using the initial model. Therefore, in the present study a statistical analysis was carried out to transform the methylation values generated by the ABI3500 capillary sequencer into values associated with the ABI3130 to make them applicable to the previously developed model. Using the transformed methylation values from samples analyzed with an ABI3500, age predictions were performed using the previously published age prediction model based on an ABI3130, which provided a MAE of ± 4.72 years.

Keywords

DNA methylation; ABI3130; ABI3500; forensic age estimation, SNaPshot; saliva; buccal cells; lineal regression

1. Introduction

Individual age estimation based on DNA methylation has become a tool of great interest and utility for the forensic field. A multitude of individual age prediction models have been generated covering different biological tissues, as well as different technologies (1). This had led to an improvement in age prediction error, but also to the identification of some constraints derived from the technology used. It has been shown that the applicability of the developed models can be limited by the analysis platform used, with the observation that samples analyzed with one platform cannot be precisely age-predicted with models built with methylation values generated from another platform. In 2018, Feng et. al (2) observed differences in prediction errors when an age prediction model generated with EpiTYPER technology was used to assessed samples analyzed with EpiTYPER (± 3.36 years) and pyrosequencing (± 4.20 years). Much larger errors were observed when attempting to directly apply methylation values generated with MPS (Massively Parallel Sequencing) to a model developed using SNaPshot (± 23.43 years) (3). In addition, with SNaPshot technology, differences were observed between different capillary electrophoresis (CE) detectors, i.e. ABI3130, ABI3500 and SeqStudio (4,5). Considering these findings and the discontinuation of older CE detectors with the release of newer instruments, existing models must be adapted, or the new data generated must be converted so that they can be correctly applied to the models developed. Several solutions have already been proposed to address this problem, such as the use of z-scores (2), the introduction of a platform variable in the age prediction models (3), or the linear regression transformation of the DNA methylation data from one technology to another (5). Considering these possibilities, as well as the discontinuation of the ABI3130, the aim of the present study was to perform linear regression to transform DNA methylation data to apply ABI3500 methylation values to the age prediction model based on the ABI3130 (6).

2. Material and Methods

In our previous study (6), we performed bisulfite conversion and SNaPshot sequencing of seven markers (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*) to generate an age prediction model based on DNA methylation values from ABI3130 analysis (Applied Biosystems, AB). To build the transformation models, 114 samples (76 buccal swabs and 38 saliva samples) from the previously analyzed samples were selected, with an age range from 22 to 84 years. To validate the transformation models, an independent set of 38 samples (18 buccal swabs and 20 saliva samples) from 23 to 80 years was used. The selected samples were analyzed simultaneously on the ABI3130 and ABI3500 (both AB). Statistical analyses were performed using R software (version 4.3.0). The Kruskal-Wallis test was used to assess whether the data generated by the different markers could be treated together or independently. Linear regression was applied to transform the DNA methylation data, and the Wilcoxon test was used to evaluate the fit of the transformed data. Statistical significance was established at a p-value of less than 0.05.

3. Results and Discussion

In total, 114 samples were evaluated to determine whether the differences observed in the methylation values obtained by both ABI3130 and ABI3500 were only due to the instrument or also marker dependent. A p-value of 2.20^{-16} was observed for the Kruskal-Wallis test, confirming that the differences between markers were significant and building individual transformation models for each marker was necessary. These differences can be seen in the scatter plots shown in Fig. 1, as well as the disparity between the values generated by the ABI3130 and ABI3500, with the methylation values of the ABI3130 being systematically higher in almost all markers. Similar trends were observed by So et al. (5) for 4 of the 5 CpGs analyzed when evaluating the differences observed between ABI3130, ABI3500 and SeqStudio instruments. In our study, the only marker that did not follow this trend was *OTUD7A*. It is important to note that this marker presented methylation values close to zero in ABI3130 and ABI3500 for the methylated peak, with values clustering in this region of the axis in the plot.

For each of the markers of interest, linear regression models were developed to transform methylation values obtained with the ABI3500. Table 1 shows the equations generated for all the analyzed markers, as well as the corresponding R-square values. The R-square values obtained were greater than 0.9 except for *OTUD7A*, with a value of 0.35, as can be expected from the observed distribution of points in the corresponding graph in Fig. 1. To assess the suitability of the transformation, a Wilcoxon test was performed between the values obtained with the ABI3130 and the values transformed from those obtained with the ABI3500. For all markers, a p-value greater than 0.05 was obtained, indicating a good fit and acceptance of the transformation models. To evaluate the age prediction model using data generated with the ABI3500, the methylation values of 38 individuals, analyzed with the ABI3500 sequencer were transformed and their ages predicted using the previously developed model (6). The ages of the testing set were also predicted with the values generated by the ABI3130 as a reference of prediction accuracy, as shown in Fig. 2. The errors obtained with ABI3130 values gave an MAE of ± 3.04 years and with the corrected values an MAE of ± 4.72 years, representing values in broad agreement with the MAE obtained in the validation of the age prediction model (MAE ± 3.66 years). Therefore, the corrected values can be directly used on the corresponding open-access *Snipper* forensic classifier (7). These results are in line with those obtained by Feng et. al (2), a study in which the transformed data gave a prediction error (± 4.20 years) one year higher than that obtained with the validation set analyzed using the same technique as the model data (± 3.36 years). Although the transformation methods are different (z-score and linear regression), the errors observed are similar. Using the same approach, So et. al (5) obtained differences of less than one year between the prediction errors of the validation sets analyzed with the technique used to develop the model and the values corrected from the two other technologies they evaluated.

4. Concluding remarks

The need to adapt current age prediction models as technologies move forward has been highlighted in several studies (2–5). However, it has also been shown that the discontinuation of CE instruments does not need to affect the models developed on these platforms. The design of transformation models (5), the use of additional variables in the models (3), or the use of correction factors (2) can extend the useful life of existing age prediction models to other technologies or detection platforms. In this study, differences in DNA methylation values corresponding to the markers selected in the construction of an age prediction model were compared between the ABI3130 and ABI3500 CE detectors, and the model previously developed (6) was easily applied to samples analyzed on the ABI3500, extending not only the lifetime of the model, but also its applicability. However, this is only an adaptative step towards creating prediction models that are not limited to a specific platform, highlighting the need to create prediction models that are not equipment dependent.

Acknowledgements

Part of this work has been supported by the Ministerio de Educación, Cultura y Ciencia, Spain (PID2019–107876RB-I00) with MVL as Principal Investigator.

References

1. Naue J. Getting the chronological age out of DNA: using insights of age-dependent DNA methylation for forensic DNA applications. Genes and Genomics. 2023;45:1239–61.

2. Feng L, Peng F, Li S, Jiang L, Sun H, Ji A, et al. Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. Forensic Sci Int Genet. 2018;35:38–46.

3. Hong S, Shin K, Jung S, Lee E, Lee H. Platform-independent models for age prediction using DNA methylation data. Forensic Sci Int Genet. 2019;38:39–47.

4. Lee J, Lee J, Naue J, Fleckhaus J, Freire-Aradas A, Neubauer J, et al. A collaborative exercise on DNA methylation-based age prediction and body fluid typing. Forensic Sci Int Genet. 2022;57:102656.

5. So MH, Lee HY. Genetic analyzer-dependent DNA methylation detection and its application to existing age prediction models. Electrophoresis. 2021;42(14–15):1497–506.

6. Ambroa-Conde A, Girón-Santamaría L, Mosquera-Miguel A, Phillips C, Casares de Cal MA, Gómez-Tato A, et al. Epigenetic age estimation in saliva and in buccal cells. Forensic Sci Int Genet. 2022;61(August).

7. Snipper App suite [Internet]. [cited 2023 Nov 20]. Available from: http://mathgene.usc.es/snipper/

Figure legends.

Figure 1. Dispersion diagrams for cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD* representing the DNA methylation values obtained with the ABI3130 in front of the corresponding values obtained using the ABI3500 for 114 individuals from 22 to 84 years old. The red continuous line represents the best fit line of the plotted data. The dotted blue line represents perfect correlation.

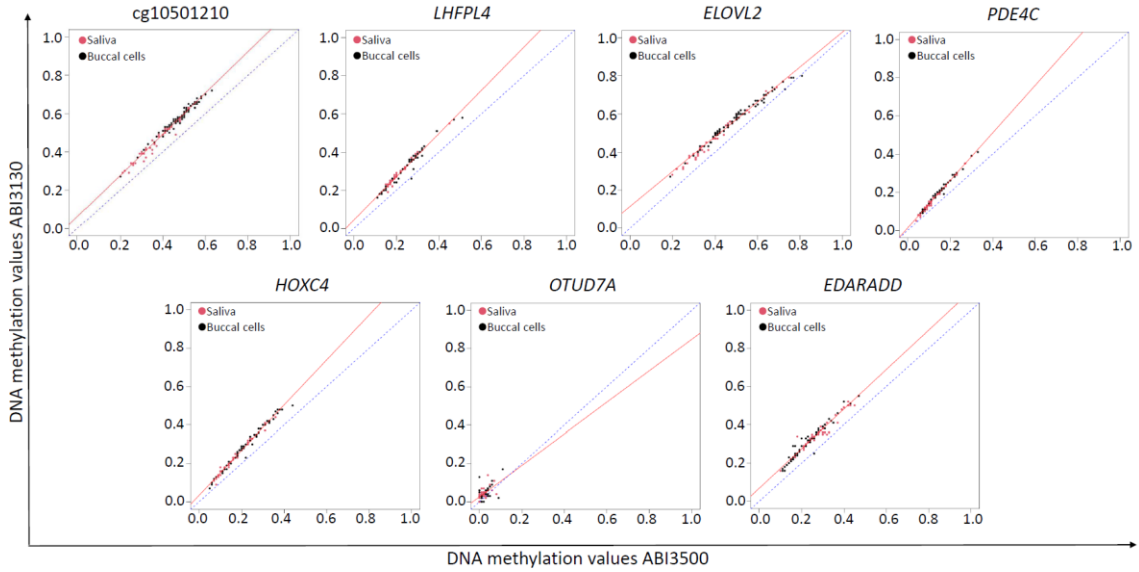


Figure 2. Dispersion diagram of the age predicted based on the ABI3130 DNA methylation values in front of the age predicted based on the transformed data generated from the DNA methylation values generated with the ABI3500 genetic analyzer. The red continuous line represents the best fit line of the plotted data. The dotted blue line represents perfect correlation.

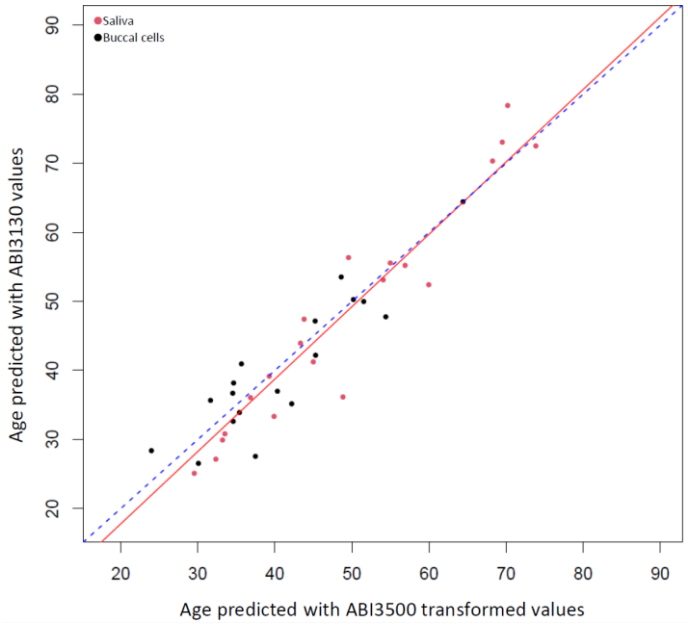


Table 1. Linear regression equations for DNA methylation correction and R-squared values for the generated transformation models generated with ABI3130 and ABI3500 methylation values.

Gene	**CpG_ID**	**Equation**[a)]	**R^2**
none	cg10501210	y = 0.06118 + 1.07840x	0.9730
LHFPL4	cg11084334	y = 0.04413 + 1.13238x	0.9613
ELOVL2	cg16867657	y = 0.1139 + 0.9156x	0.9794
PDE4C	none	y = 0.01282 + 1.24526x	0.9858
HOXC4	cg18473521	y = 0.03381 + 1.17573x	0.9827
OTUD7A	cg04875128	y = 0.02541 + 0.82235x	0.3531
EDARADD	cg09809672	y = 0.06831 + 1.03410x	0.9490

a) y = DNA methylation value from ABI3130; x = DNA methylation value from ABI3500

ASPECTOS ÉTICOS

ASPECTOS ÉTICOS

Conflicto de interés:

Adrián Ambroa Conde, como autor de esta tesis doctoral, declara que no hay conflictos de interés asociados a este trabajo.

Aprobaciones de comités de ética:

A continuación, se adjuntan los documentos emitidos por el Comité Autonómico de Ética de la Investigación en Galicia (CAEI) y el Comité Bioético de Jagiellonian de la universidad de Krak´ow en Polonia (KBET), por las que se aprueban la realización de las investigaciones asociadas al proyecto Epigenética: Desarrollo de nuevas aplicaciones en genética forense (CAEI: 2013/543 y KBET/122.6120.86.2017).

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral

Comité Autonómico de Ética de la Investigación de Galicia
Edificio Administrativo de San Lázaro
15781 SANTIAGO DE COMPOSTELA
Tlf: 881 546425 Fax: 881 541804
ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ AUTONÓMICO DE ÉTICA DE LA INVESTIGACIÓN DE GALICIA

Paula M. López Vázquez, Secretaria del Comité Autonómico de Ética de la Investigación de Galicia

CERTIFICA:

Que este Comité evaluó en su reunión del día 28/01/2014 el estudio:

Título: Desarrollo de técnicas cronométricas utilizando marcadores epigenéticos para el análisis en Genética Forense-CRONOGEN
Promotor: María Victoria Lareu Huidobro
Código de Registro CAEI de Galicia: 2013/543

Y, tomando en consideración las siguientes cuestiones:

- La pertinencia del estudio, teniendo en cuenta el conocimiento disponible, así como los requisitos legales aplicables, y en particular la Ley 14/2007, de investigación biomédica, el Real Decreto 1716/2011, de 18 de noviembre, por el que se establecen los requisitos básicos de autorización y funcionamiento de los biobancos con fines de investigación biomédica y del tratamiento de las muestras biológicas de origen humano, y se regula el funcionamiento y organización del Registro Nacional de Biobancos para investigación biomédica, la ORDEN SAS/3470/2009, de 16 de diciembre, por la que se publican las Directrices sobre estudios Posautorización de Tipo Observacional para medicamentos de uso humano, y el la Circular nº 07 / 2004, investigaciones clínicas con productos sanitarios.
- La idoneidad del protocolo en relación con los objetivos del estudio, justificación de los riesgos y molestias previsibles para el sujeto, así como los beneficios esperados.
- Los principios éticos de la Declaración de Helsinki vigente.
- Los Procedimientos Normalizados de Trabajo del CEIC de Galicia

Emite un **INFORME FAVORABLE** para la realización del estudio por el/la investigador/a del centro:

Centros	Investigadores Principales
USC (Instituto de Ciencias Forenses Luis Concheiro)	María Victoria Lareu Huidobro

En Santiago de Compostela, a 31 de enero de 2014
La Secretaria

Paula M. López Vázquez

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ AUTONÓMICO DE ÉTICA DE LA INVESTIGACIÓN DE GALICIA

Paula M. López Vázquez, Secretaria del Comité Autonómico de Ética de la Investigación de Galicia,

CERTIFICA:

Que este Comité evaluó en su reunión del día 24/01/17, la enmienda del estudio:

Título: Desarrollo de técnicas cronométricas utilizando marcadores epigenéticos para el análisis en Genética Forense-CRONOGEN
Versión Enmienda: *Modificación de 09/12/2016 (actualización de protocolo y documentos de consentimiento informado)*
Promotor: *María Victoria Lareu Huidobro*
Código de Registro: 2013/543

Y que este Comité acepta de conformidad con sus procedimientos normalizados de trabajo y tomando en cuenta los requisitos éticos, metodológicos y legales exigibles a los estudios de investigación con seres humanos, sus muestras o registros, que dicha enmienda sea incorporada al estudio de investigación que se está realizando en los centros aprobados.

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

galicia

Y HACE CONSTAR QUE:

1 El CAEIG cumple los requisitos legales vigentes (R.D 1090/2015 y la Ley 14/2007.
2 El CAEIG tanto en su composición como en sus PNTs cumple las Normas de Buena Práctica Clínica (CPMP/ICH/135/95).
3 La composición actual del CAEIG es:

Manuel Portela Romero. (Presidente). Médico Especialista en Medicina Familiar y Comunitaria.
Irene Zarra Ferro. (Vicepresidenta). Farmacéutica de Atención Especializada.
Paula Mª López Vázquez, (Secretaria). Médico Especialista en Farmacología Clínica.
Juan Vázquez Lago (Secretario Suplente). Médico Especialista en Medicina Preventiva y Salud Pública.
Jesús Alberdi Sudupe. Médico especialista en Psiquiatría.
Rosendo Bugarín González. Médico Especialista en Medicina Familiar y Comunitaria.
Juan Casariego Rosón. Médico Especialista en Cardiología.
Xoán X. Casas Rodríguez. Médico Especialista en Medicina Familiar y Comunitaria.
Juana Mª Cruz del Río. Trabajadora Social.
Juan Fernando Cueva Bañuelos. Médico Especialista en Oncología Médica.
José Álvaro Fernández Rial. Médico Especialista en Medicina Interna.
José Luis Fernández Trisac. Médico Especialista en Pediatría.
Mª José Ferreira Díaz. Diplomada Universitaria de Enfermería
Pablo Nimo Ríos. Licenciado en Derecho.
Pilar Gayoso Diz. Médico Especialista en Medicina Familiar y Comunitaria.
Agustín Pía Morandeira. Farmacéutico de Atención Primaria
Salvador Pita Fernández. Médico Especialista en Medicina Familiar y Comunitaria.
Carmen Rodríguez-Tenreiro Sánchez. Licenciada en Farmacia.
Susana María Romero Yuste. Médico Especialista en Reumatología.
Mª Asunción Verdejo González. Médico Especialista en Farmacología Clínica.

En Santiago de Compostela, a 26 de enero de 2017

XUNTA DE GALICIA SECRETARÍA XERAL TÉCNICA COMITÉ AUTONÓMICO DE ÉTICA DA INVESTIGACIÓN CONSELLERÍA DE SANIDADE

Firmado digitalmente por PAULA MARÍA LÓPEZ VÁZQUEZ - 46900339G
Nombre de reconocimiento (DN): 2.5.4.13=Qualified Certificate: AAPP-FP-M-HW-KUSU, cn=PAULA MARÍA LÓPEZ VÁZQUEZ - 46900339G, givenName=PAULA MARÍA, sn=LÓPEZ VÁZQUEZ, serialNumber=46900339G, ou=certificado electrónico de empleado público, o=XUNTA DE GALICIA, c=ES
Fecha: 2017.01.26 08:47:06 +01'00'

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIWERSYTET JAGIELLOŃSKI W KRAKOWIE

OPINIA
nr 122.6120.86.2017 z dnia 28 kwietnia 2017 roku

Na zebraniu w dniu 28 kwietnia 2017 r. Komisja zapoznała się z wnioskiem z dnia 12 kwietnia 2017 r.
złożonym:
przez kierownika tematu: **dr n. med. Danuta Piniewska - Róg**
zatrudnionego **Katedra i Zakład Medycyny Sądowej UJCM**
31 – 531 Kraków, ul. Grzegórzecka 16

oraz jego merytorycznym uzasadnieniem dotyczącym przeprowadzenia eksperymentu medycznego pt. „Walidacja biomarkerów wieku człowieka dla celów sądowo – lekarskich – projekt VISAGE".

Komisja Bioetyczna
Uniwersytetu
Jagiellońskiego

Do wniosku dołączono:

1. Oświadczenie o braku załączenia formularza informacji dla pacjenta, formularza zgody uczestnika badania, formularza o ochronie danych osobowych, wersja 1.
2. Oświadczenie Wnioskodawcy o statusie prawnym dotyczącym wykorzystania próbek materiału sekcyjnego z rutynowych badań sądowo – lekarskich do projektu badawczego oraz ochrony danych o zwłokach, wersja 1 z dnia 10.04.2017 r.
3. Życiorys naukowy Wnioskodawcy, wersja 1.
4. Lista najważniejszych publikacji, wersja 1 z dnia 10.04.2017 r.
5. Protokół badania, wersja 1 z dnia 10.04.2017 r.
6. Oświadczenie o realizacji projektu w ramach prac badawczych UJ/UJCM.

Komisja wyraża pozytywną opinię w sprawie przeprowadzenia wnioskowanego badania - na warunkach określonych we wniosku oraz dodatkowo zastrzegając:
1/ obowiązek przedstawienia Komisji:
- wszystkich zmian w protokole mających wpływ na przebieg oraz ocenę badania,
- zawiadomienia o przyczynach przedwczesnego zakończenia badania,
- sprawozdania w toku przeprowadzanych badań - co sześć miesięcy,
- raportu końcowego.

Badanie może być prowadzone do dnia 28 kwietnia 2018 roku.
Skład i działanie Komisji zgodne z GCP oraz wymogami lokalnymi. Lista członków Komisji biorących udział w posiedzeniu stanowi załącznik do niniejszego dokumentu.

Kraków, dnia 28 kwietnia 2017 r.

Przewodniczący
Komisji Bioetycznej UJ
prof. dr hab. n. med. Piotr Thor

ul. Podwale 3/5
PL 31-118 Kraków
tel. +48 (12) 37 04 386
kbet@cm-uj.krakow.pl
www.kbet.cm-uj.krakow.pl

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

AKCEPTACJA

dot. opinii nr: **122.6120.86.2017z dnia 28 kwietnia 2018 roku**

TYTUŁ BADANIA:
„Walidacja biomarkerów wieku człowieka dla celów sądowo-lekarskich – projekt VISAGE"

WNIOSKODAWCA:
dr n. med. Danuta Piniewska-Róg
Katedra i Zakład Medycyny Sądowej UJ CM
31-531 Kraków, ul. Grzegórzecka 16

Komisja Bioetyczna
Uniwersytetu
Jagiellońskiego

PRZEDSTAWIONE DOKUMENTY:
1) Zgłoszenie poprawki z dnia 05.11.2019 r.;
2) Wystąpienie o brak sprzeciwu prokuratora.

Komisja Bioetyczna Uniwersytetu Jagiellońskiego na posiedzeniu w dniu 21 listopada 2019 r., po zapoznaniu się z wyżej wymienionymi dokumentami pozytywnie zaopiniowała zgłoszoną poprawkę oraz wyraziła zgodę na **wznowienie badania do 30 kwietnia 2021 roku.**

Lista członków Komisji Bioetycznej biorących udział w posiedzeniu:

Przewodniczący: prof. dr hab. med. Piotr Thor – lekarz – chirurg ogólny/urolog
Zastępca przewodniczącego: mgr Alicja Widera – psycholog kliniczny

Członkowie:
prof. dr hab. med. Roman Pfitzner – lekarz – chirurg ogólny/kardiochirurg
dr hab. med. Ewa Konduracka, prof. UJ – lekarz – internista/kardiolog
dr hab. med. Klaudia Stangel-Wójcikiewicz – lekarz – ginekolog-położnik
dr hab. n. med. Tomasz Kaczmarzyk, prof. UJ – lekarz stomatolog – chirurg stomatolog
dr hab. med. Piotr Major, prof. UJ – lekarz – chirurg ogólny
dr med. Stefan Bednarz – lekarz – internista – przedstawiciel Okręgowej Rady Lekarskiej w Krakowie
mgr Leszek Kądziela – radca prawny

Skład i działanie Komisji zgodne z GCP oraz wymogami lokalnymi

Kraków, 21 listopada 2019 r.

Przewodniczący
Komisji Bioetycznej UJ

prof. dr hab. n. med. Piotr Thor

ul. Grzegórzecka 20
PL 31-531 Kraków
tel. + 48 (12) 433 27 39
+ 48 (12) 433 27 43
kbet@cm-uj.krakow.pl
www.kbet.cm-uj.krakow.pl

La metilación del ADN es un biomarcador epigenético que ha despertado un gran interés en el campo forense en los últimos años. Los objetivos de la presente tesis doctoral han consistido en la generación de modelos estadísticos basados en este biomarcador, que puedan ser empleados como herramientas moleculares con el fin de orientar determinadas investigaciones policiales. Con esto en mente se han desarrollado modelos para la estimación de la edad cronológica, la identificación de tejidos y la inferencia de estilos de vida, en concreto consumo de tabaco y alcohol.