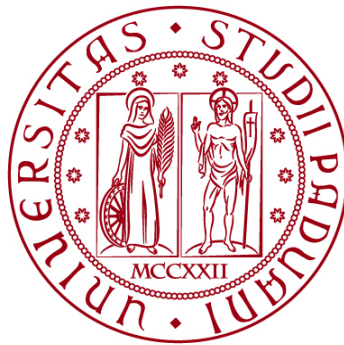


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia Molecolare



ELABORATO DI LAUREA

**SVILUPPO DI KIT PREDITTIVI PER L'ABITUDINE AL FUMO A
PARTIRE DALL'ANALISI DEL PATTERN DI METILAZIONE DI
13 SITI CpG**

**Tutor: Prof. Giovanni Vazza
Dipartimento di Biologia**

Laureanda: Giulia Alvaro

ANNO ACCADEMICO 2022/2023

Sommario

CAPITOLO 1: INTRODUZIONE	1
1.1 Tipi di marcatori utilizzati nelle analisi forensi	1
1.2 Epigenetica: qual è il ruolo dell'epigenetica in ambito forense e cosa si intende per metilazione del DNA.	1
1.3 Fumo: dati e rilevanza in ambito forense.....	3
1.4 Fumo e predittori molecolari	3
CAPITOLO 2: MATERIALI E METODI	4
2.1 Campioni, controlli e preparazione dei campioni di DNA.....	4
2.2 Sequenziamento diretto (BSP), preparazione e quantificazione di librerie senza PCR	5
2.3 MPS basato su Illumina, analisi dei dati MPS e modellizzazione della previsione	6
CAPITOLO 3: RISULTATI	7
3.1 Valutazione dei bias di amplificazione dei campioni target utilizzando controlli di DNA metilato artificialmente.	8
3.2 Assegnazione dei campioni in studio ai fenotipi del fumo utilizzando i modelli di previsione.....	9
3.3 Caratteristiche aggiuntive del pool di marcatori delle 13CpG.....	10
3.4 Previsione dell'abitudine al fumo utilizzando modelli statistici vecchi e nuovi.	11
CAPITOLO 4: DISCUSSIONE	12
4.1 Nuovo modello basato su MPS e i suoi limiti	13
4.2 Scelta delle CpG e limiti di classificazione	14
4.3 Altri fattori che influiscono sull'analisi condotta	15
4.4 Pattern di metilazione e stile di vita	16
CONCLUSIONI	16

ABSTRACT

La predizione dell'abitudine al fumo a partire dall'analisi epigenetica, è considerata una nuova frontiera in ambito forense, usata per ottenere una fenotipizzazione più accurata del DNA. In questo studio, si è cercato di sviluppare kit predittivi per classificare l'abitudine al fumo, basandosi sulla metilazione di 13CpG del sangue. L'analisi, dopo il prelievo dei campioni, l'estrazione del DNA genomico e la conversione con bisolfito, è passata tramite una multiplex PCR, preparazione di librerie prive di amplificazione e MPS mirato paired-end. Dopo aver analizzato e corretto eventuali bias riscontrati nell'amplificazione, tramite modelli congiunti con correzioni inter-tecnologiche, i risultati ottenuti dall'applicazione sia del modello a due categorie (fumatori attuali/ex) che di quello a tre categorie (mai fumatori/ex/attuali), su uno strumento MPS a 232 campioni di sangue europei, ha fornito risultati predittivi migliori. Si è giunti alla conclusione che 11 su 13CpG del fumo sono correlate alle sigarette giornaliere nei fumatori attuali, solo una è correlata debolmente al tempo trascorso dalla cessazione negli ex fumatori ed 8 CpG sono correlate all'età, mentre una mostra differenze sesso-dipendenti. I modelli sviluppati non risultano comunque essere definitivi, ma servono nuove ricerche per validare i test dal punto di vista di sensibilità e sui biomarcatori da utilizzare.

CAPITOLO 1: INTRODUZIONE

1.1 Tipi di marcatori utilizzati nelle analisi forensi

Fino ad ora, nelle analisi forensi, sono stati utilizzati come marcatori gli STRs (Short Tandem Repeats), corte regioni di DNA ripetute, altamente polimorfe e facili da amplificare tramite PCR (Polymerase Chain Reaction). Il polimorfismo risiede nella lunghezza della regione ripetuta; inoltre, l'identità dell'individuo corrispondente al campione prelevato dalla scena del crimine deriva da comparazione del profilo ottenuto con profili presenti all'interno di database delle forze dell'ordine o tramite confronto con profili di persone sospettate. Si tratta di un metodo con elevata capacità discriminatoria, che permette di utilizzare una piccola quantità di DNA, anche di scarsa qualità, ma che non consente di ottenere informazioni aggiuntive nel momento in cui non si trova corrispondenza tra campione prelevato e profili disponibili. Per superare questa problematica, si possono utilizzare degli approcci innovativi basati sulla genotipizzazione di specifici SNPs (Single Nucleotide Polymorphisms) con lo scopo di ottenere informazioni sul fenotipo del soggetto donatore: si parla di FDP (Forensic DNA Phenotyping) riferendosi alla possibilità di previsione di tratti dell'aspetto sconosciuti, non identificabili con profilo comparativo del DNA, a partire da un campione biologico derivante da donatori o persone decedute di cui non si conosce l'identità. A questo scopo, è possibile utilizzare VISAGE Enhanced Tool for Appearance and Ancestry, primo strumento sviluppato a scopi forensi che include marcatori consolidati per la predizione del colore di occhi, capelli, pelle, sopracciglia, calvizie maschili, presenza di lentiggini e discendenza biogeografica (quest'ultima comprende SNPs autosomici). VISAGE Enhanced Tool for Appearance and Ancestry è un test MPS (Massive Parallel Sequencing) robusto, riproducibile e, per il grande numero di SNPs, abbastanza sensibile, con alti tassi di concordanza.

1.2 Epigenetica: qual è il ruolo dell'epigenetica in ambito forense e cosa si intende per metilazione del DNA.

In ambito forense, l'epigenetica è un campo esplorato solo recentemente, in particolare in questo ambito il biomarcatore più utilizzato è la metilazione. La metilazione del DNA è la modifica epigenetica più frequente che consiste nell'aggiunta di un gruppo metile, derivante dall'S-adenosilmetionina, al carbonio 5 della citosina, dando vita alla 5-metilcitosina (5mC). La modifica non è randomica, infatti si riscontra a livello delle isole CpG (regioni genomiche ben definite in cui la frequenza del dinucleotide CG è maggiore rispetto al caso), che troviamo per lo più a livello promotoriale e che hanno quindi molta rilevanza nella

regolazione dell'espressione genica. La metilazione delle isole CpG risulta essere correlata a una modifica della cromatina, in particolare, alla deacetilazione a livello dell'istone H3: i doppietti CG metilati sono legati da MBD proteins (Methyl Binding Domain proteins), come MeCP1, le quali a loro volta interagiscono con complessi enzimatici che deacetilano la cromatina, causando la sua compattazione con conseguente repressione dell'attività trascrizionale. A livello della cellula differenziata, i geni posseggono un pattern di metilazione del DNA stabile e tessuto-specifico, da cui ne deriva una trascrizione tessuto-specifica; nell'adulto, la maggior parte delle isole CpG presenti nelle regioni promotoriali risulta essere ipometilata e ciò corrisponde ad attività trascrizionale, mentre la maggior parte delle CpG fortemente metilate si trovano a livello delle regioni inter-geniche. Una stretta correlazione tra metilazione e regolazione dell'attività trascrizionale, risulta essere rilevante nei geni in cui si identificano le "weak CpG island" o "CpG island shores", regioni dove la percentuale dei dinucleotidi CG si aggira attorno al 30%, e la cui metilazione risulta essere dinamica e tessuto-specifica. L'introduzione di nuovi pattern di metilazione, sia in ambito fisiologico che patologico, avviene a opera delle metilasi de novo DNMT3A e DNMT3B, mentre il mantenimento del pattern di metilazione avviene a opera della metilasi di mantenimento DNMT1. Attualmente sono emerse prove convincenti del fatto che anche i lncRNA, sono coinvolti nella determinazione del pattern di metilazione de novo e del mantenimento del pattern di metilazione, sia in condizioni fisiologiche che patologiche, in quanto sono in grado di reclutare DNMT3A, DNMT3B e DNMT1, in maniera diretta ma anche indiretta, influenzando così nella regolazione di geni bersaglio, che risultano avere ruoli chiave nell'impegno del mesoderma, nella rigenerazione muscolare, nella differenziazione neurale, nell'adipogenesi, nei disturbi mentali, nelle malattie cardiovascolari, nell'osteoartrite e in alcuni tipi di cancro. In particolare, l'importante ruolo della metilasi DNMT3A è sottolineato dal fatto che mutazioni a livello di questa metilasi sono collegate all'insorgenza di leucemia e di tumori ematologici.

Ad oggi, l'analisi di marcatori CpG è stata impiegata in ambito forense per l'identificazione di fluidi e tessuti corporei, per la previsione dell'età cronologica e la discriminazione di gemelli monozigoti; tuttavia, negli ultimi tempi, si è valutata la possibilità di predire, tramite il profilo epigenomico, tratti delle abitudini di vita utili a fornire un FDP più ampio.

1.3 Fumo: dati e rilevanza in ambito forense

Il fumo di tabacco è un'abitudine molto diffusa: si stima che il 22.3% della popolazione mondiale fuma, di cui il 36.7% è rappresentata da uomini e il 7.8% è rappresentato da donne. Si tratta di più dell'80% della popolazione che vive in Paesi a basso e medio reddito; in particolare, il 5-30% della popolazione europea, fuma in maniera abitudinaria. Il fumo è responsabile dell'insorgenza di vari tipi di tumore, malattie cardiovascolari, accorciamento dei telomeri, perdita di pluripotenza e, in base al tempo di esposizione al fumo, apoptosi delle cellule embrionali staminali, ipossia ed alterazioni epigenetiche. A livello epigenetico, il fumo influenza l'espressione delle DNMT: la nicotina regola negativamente DNMT1, in particolare a livello dei neuroni GABAergici, causando una riduzione dei livelli di metilazione delle isole CpG fondamentali per il corretto funzionamento dei recettori nicotinici; inoltre il condensato del fumo di sigaretta aumenta l'espressione di Sp1, fattore di trascrizione che lega i motivi CG rich nei promotori genici, impedendo la metilazione de novo. La pre-determinazione dell'abitudine del fumo risulta essere rilevante, non solo in ambito forense, ma anche nei campi di salute pubblica e medicina personalizzata per la convalida di cartelle cliniche elettroniche o questionari di partecipanti ad una ricerca, i quali spesso contengono risultati errati o mancanti. Una soluzione possibile è l'analisi tossicologica che permette di riscontrare la presenza di nicotina o suoi metaboliti (in particolare la coitina) nel sangue, ma tale analisi ha costi elevati e bassa specificità a causa della breve emivita dei metaboliti nicotinici nel sangue. Per tali motivi si è considerata la possibilità di predire l'abitudine al fumo utilizzando analisi epigenetiche.

1.4 Fumo e predittori molecolari

Negli ultimi anni, tramite analisi *genome-wide*, cioè un'analisi svolta sull'intero genoma, sono state identificate numerose isole CpG in geni diversi il cui stato di metilazione è associato al fumo. Ne sono alcuni esempi le CpG del recettore 3 della trombina o della tripsina, della fosfatasi alcalina, del recettore 15 accoppiato a proteina G e della miosina 1. I cambiamenti nella metilazione delle singole CpG indotti dal fumo, che sono stati riscontrati, sono costanti ma relativamente piccoli (mediamente al di sotto del 20%) e generalmente corrispondono ad una diminuzione dei livelli di metilazione [1]. La costruzione di predittori molecolari del fumo parte dall'analisi del profilo di metilazione del DNA genomico estratto da campioni di sangue, dall'utilizzo di piattaforme Microarray-Illumina per il rilevamento del pattern di metilazione e dall'utilizzo di metodi statistici utili a correlare il pattern di metilazione ad una classe fenotipica, che può corrispondere a fumatore, ex fumatore o mai fumatore. Studi condotti in precedenza si erano proposti di costruire predittori a due o tre categorie per l'abitudine al fumo. Lo

studio di Sugden et al.[2] su 2623 CpG correlate al fumo, ha permesso di capire come il fumo influenza la metilazione del DNA, indipendentemente dai fattori di rischio genetici e ambientali, e inoltre, ha permesso di capire come questi cambiamenti legati al fumo abbiano effetto su cambiamenti nell'espressione genica in percorsi correlati a infiammazione, risposta immunitaria e traffico cellulare; in aggiunta, tramite il calcolo di un punteggio standardizzato sulle 2623 CpG analizzate, è stato possibile discriminare con successo fumatori attuali, ex fumatori e mai fumatori. McCartney et al. [3] hanno proposto un modello statistico basato sull'analisi di 233 CpG, riuscendo a discriminare con estrema precisione fumatori attuali e mai fumatori. Bollepalli et al. [4] hanno costruito un modello basato su 121 CpG per la previsione di fumatori attuali, mai fumatori ed ex fumatori, testato in tre set indipendenti: fumatori attuali e mai fumatori sono stati individuati rispettivamente con sensibilità dell'81%-94% e specificità dell'85%-57%; per gli ex fumatori il modello ha mostrato bassa sensibilità (18%) ma specificità del 96%. Nello studio analizzato in questo elaborato, invece, la previsione dell'abitudine al fumo, è stata fatta tramite l'analisi, con tecnologia MPS, di 13 CpG fortemente associate al fumo, selezionate in uno studio precedente, da Maas et al. [5]. Nello studio di Maas, i ricercatori sono partiti dall'analisi di sei coorti di soggetti olandesi (n= 3764), a cui sono stati aggiunti altri 646 partecipanti, con lo scopo di costruire predittori a due e tre categorie, basandosi sull'analisi di sole 13 CpG del fumo tramite tecnologia Microarray. Le CpG sono state selezionate in base al fatto che le stesse CpG fossero state evidenziate in almeno due studi, e in cui, i livelli di metilazione mostrassero almeno il 10% di differenza, e stessa direzione in tutti gli studi esaminati.

CAPITOLO 2: MATERIALI E METODI

2.1 Campioni, controlli e preparazione dei campioni di DNA

Lo studio è stato condotto a partire dalla raccolta dei campioni da analizzare: sono stati selezionati 232 europei residenti nei Paesi Bassi, di cui 108 maschi e 124 femmine di età che va dai 22 agli 82 anni, inseriti nelle categorie di fumatori attuali, ex fumatori e mai fumatori in base ai dati raccolti tramite questionari autoriferiti; in particolare, dei 232 partecipanti, 90 sono fumatori attuali, 71 sono ex fumatori e 71 sono mai fumatori. È stato raccolto poi un campione di sangue venoso periferico per ogni partecipante e sono stati preparati 9 campioni artificialmente metilati in diversi rapporti (0, 10, 20, 40, 50, 60, 80, 90, 100%), ottenuti mescolando DNA umano di controllo metilato e DNA umano di controllo non metilato. Successivamente è stato estratto il DNA, è stato quantificato, è stata normalizzata la concentrazione di DNA a 4ng/μL, è stato incubato per 16 ore con lo scopo di raggiungere la massima conversione in bisolfito e, in seguito, 200ng di

campione di DNA genomico sono stati sottoposti a trattamento con bisolfito. La conversione del DNA con bisolfito ha lo scopo di determinare il pattern di metilazione del DNA genomico: il bisolfito converte i residui di citosina non metilati in uracile e lascia i residui di 5-metilcitosina inalterati. In questo modo vengono introdotti dei cambiamenti nella sequenza di DNA, dipendenti dallo stato di metilazione dei singoli residui di citosina, e questo ci permette di ottenere informazioni sullo stato di metilazione del DNA a livello di singolo nucleotide. La concentrazione del DNA convertito con bisolfito e l'efficienza di conversione sono state valutate in duplicato tramite real-time PCR. Infine è stata normalizzata la concentrazione di DNA convertito in bisolfito a 10ng/μL.

2.2 Sequenziamento diretto (BSP), preparazione e quantificazione di librerie senza PCR

L'analisi del DNA convertito con bisolfito è stata eseguita tramite PCR seguita da un sequenziamento Illumina. Questo tipo di analisi prevede la costruzione di primers che si appaiano al di fuori delle CpG di interesse; si tratta di primers che non sono sovrapposti ai siti di metilazione indagati ma che li fiancheggiano. In questo studio, sono state analizzate 13 CpG del fumo selezionate precedentemente nel lavoro di Maas e collaboratori [5]. Per ottenere la sequenza genomica fiancheggiante le 13 CpG usate come marcatori, è stato utilizzato il genome browser Ensembl. Per estrarre le sequenze genomiche convertite con bisolfito, i primers sono prima stati progettati tramite il software MethPrimer e successivamente testati tramite il software BiSearch, il quale, tramite PCR elettronica (ePCR), testa i primers disegnati per rilevare potenziali siti di innesco e prodotti di PCR indesiderati. Il DNA trattato con bisolfito, risulta essere difficile da amplificare a causa della ridondanza della sequenza di DNA ottenuta dopo la conversione (sequenza ricca in T); per questo motivo, si può riscontrare una bassa efficienza di amplificazione e presenza di prodotti di PCR aspecifici. Per far fronte a queste problematiche, il controllo dei primers progettati con BiSearch risulta essere fondamentale al fine di ottenere dati attendibili. Allo stesso scopo è stato utilizzato il software Autodimer, il quale utilizza un algoritmo per prevedere la potenziale formazione di dimeri o forcine intramolecolare tra i primers, fornendo come output un punteggio che indica il grado di interazione. Sono stati poi condotti diversi saggi di PCR single-plex e successivamente una gel elettroforesi per valutare la giusta temperatura di annealing, la giusta concentrazione dei primers e di MgCl₂; quindi, in base alla temperatura di annealing condivisa tra i primers progettati, le 13 CpG del fumo sono state suddivise in 3 multiplex PCR separate. Dopo aver allestito la reazione di PCR, i prodotti della reazione sono stati raggruppati e purificati per poi essere quantificati in duplicato: le misure fatte sulle repliche sono state utili per normalizzare la concentrazione di DNA a 10ng/μL.

Parte degli ampliconi, ottenuti dal saggio di PCR (500ng), sono stati purificati e utilizzati per preparare librerie prive di PCR, con lo scopo di evitare ulteriori bias di amplificazione, derivanti dal fatto che l'amplificazione risulta essere preferenziale per gli alleli non metilati. Il DNA non è stato ulteriormente frammentato perché dalla PCR sono stati generati ampliconi di dimensioni adeguate (minore di 300 paia di basi). Le estremità degli ampliconi sono state rese blunt e infine adenilate per poi legarci degli adattatori contenenti gli indici e aventi le estremità T sporgenti. Le librerie sono state successivamente purificate e quantificate. Infine, le librerie indicizzate, sono state messe insieme ad una concentrazione finale di 2nM in 300 μ L di volume per poi essere riquantificate tramite real-time PCR per determinare il volume ottimale da caricare nel sequenziatore.

2.3 MPS basato su Illumina, analisi dei dati MPS e modellizzazione della previsione

Le librerie sono state denaturate in NaOH e poi diluite in un tampone di ibridazione per ottenere libreria da 8-9pM; quindi è stata aggiunta la libreria di controllo PhiX, la quale, avendo una composizione nucleotidica varia (45% GC e 55% AT) fornisce ad ogni ciclo di sequenziamento segnali fluorescenti bilanciati che mancano in librerie di campioni a bassa diversità, col fine di migliorare la qualità generale della corsa. È stato fatto poi un sequenziamento paired-end. L'analisi dei dati è stata condotta tramite un tool informatico (GAMBA), al quale vengono dati in input un insieme di file FASTAQ di lettura paired-end: GAMBA taglia le reads in base alle sequenze dei primers e le filtra in base alla loro dimensione e ad un punteggio di qualità predefinito. Prima o dopo l'elaborazione dei dati, si fa un controllo qualità tramite FastQC, il quale fornisce una rapida panoramica per individuare le aree in cui potrebbero esserci problemi e fornisce grafici e tabelle riepilogative per valutare rapidamente i dati. Per tutte le coppie di reads, GAMBA allinea le reads convertite in bisolfito, generando una libreria non direzionata, la quale è poi analizzata tramite il software Bismark, che esegue sia mappatura, che lettura, che "chiamata di metilazione" riuscendo a discriminare citosine nel contesto CpG rispetto alle altre, con lo scopo di permettere di visualizzare e interpretare i dati di metilazione subito dopo il sequenziamento. GAMBA combina i conteggi delle reads di citosine metilate e non metilate tra i vari campioni e poi ordina e indicizza i file BAM per inserirle nell'Interactive Genomics Viewer (IGV), strumento interattivo utile per l'esplorazione visiva dei dati genomici. L'elaborazione e la visualizzazione generale dei dati è stata eseguita nel linguaggio di programmazione R, un software per calcolo statistico e grafica. Le prestazioni del metodo di conversione in bisolfito, con successiva multiplex PCR, preparazione di librerie senza PCR e paired-end MPS su Illumina, sono state valutate utilizzando

duplicati tecnici (*Figura 1*). Questo è stato utile per comprendere i bias di amplificazione, e correggerli tramite un modello statistico (bi-esponenziale), calibrando il modello tramite i dati derivanti dall'analisi del DNA metilato artificialmente, il quale fornisce risultati privi di bias.

CAPITOLO 3: RISULTATI

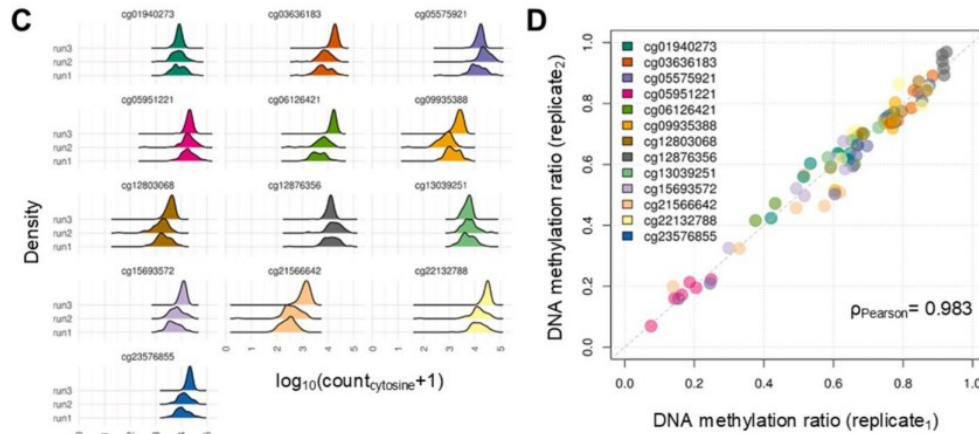


Figura 1: (C) si possono osservare diagrammi di densità di cresta che mostrano il coverage di lettura di ogni CpG per ogni corsa; (D) osserviamo un grafico di regressione lineare che mostra l'accordo nella metilazione del DNA rilevata (senza correzione del bias della PCR) per sei duplicati tecnici tra CpG.

Per produrre dati di metilazione robusti e riproducibili per ogni campione in studio sono state ottenute 30 milioni di reads. Le reads medie per ogni campione variano in modo significativo: si ha un intervallo di reads che va dalle 740 alle 21.756, con 2 CpG in particolare (cg0595122 e cg22132788) molto performanti e altre 2 CpG (cg09935388 e cg21566642) che offrono le peggiori prestazioni. L'efficienza della conversione risulta essere maggiore del 99% e questo dato deriva dal fatto che gli ampliconi convertiti in bisolfito hanno in media 50 citosine non metilate e non appartenenti alle CpG. Si può affermare che si ha un'amplificazione selettiva dei filamenti di DNA convertiti. Come abbiamo visto in *Figura 1*, per valutare la riproducibilità della quantificazione della metilazione basata su saggio MPS sono stati utilizzati sei duplicati tecnici di campioni scelti casualmente dal set di campioni ematici, analizzati separatamente su diverse corse MPS. Facendo sempre riferimento alla *Figura 1*, l'analisi delle CpG contenute nei sei duplicati mostrano un accordo nella metilazione molto elevato.

3.1 Valutazione dei bias di amplificazione dei campioni target utilizzando controlli di DNA metilato artificialmente.

È noto che i saggi di PCR con bisolfito offrano tassi di amplificazione diversi tra ampliconi metilati e non, a causa delle ampie differenze di sequenza dovute alla conversione differenziale dei siti CpG, con conseguente impatto negativo nella quantificazione della metilazione. Per risolvere questa problematica, sono stati utilizzati dei controlli di DNA metilato artificialmente (che chiamiamo standard), in varie percentuali, analizzati in duplicato in 2 corse MPS differenti.

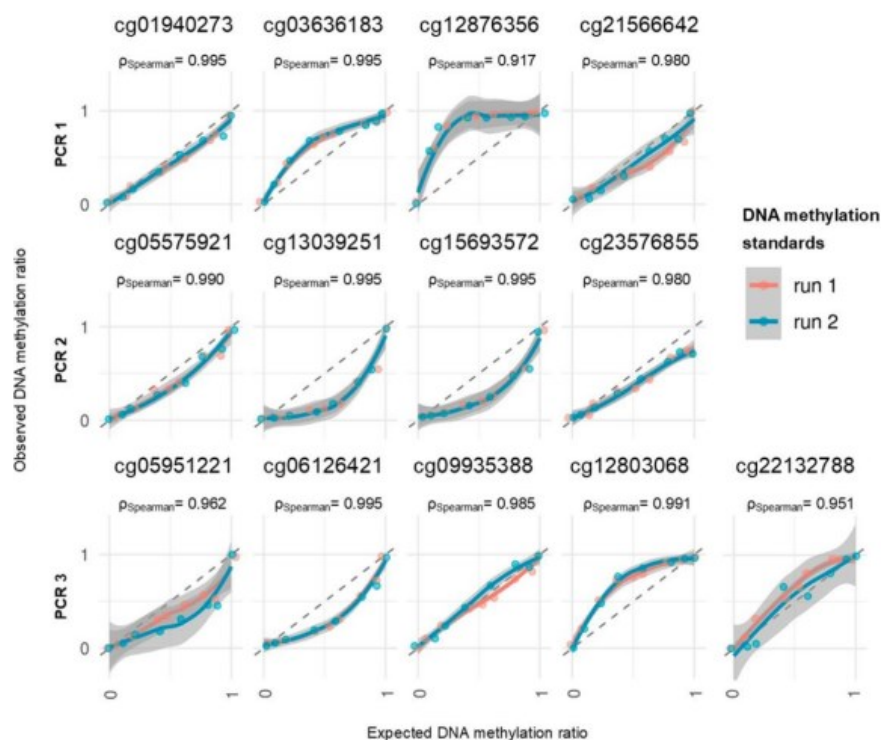


Figura 2: curve di calibrazione ottenute analizzando standard di DNA metilato artificialmente (0, 10, 20, 40, 50, 60, 80, 90, 100 %) per ognuna delle 13 CpG in esame, nelle due corse MPS.

Il grafico della metilazione osservata rispetto a quella attesa (Figura 2), in tutti i target CpG, ha evidenziato una deviazione rispetto al rilevamento lineare; in particolare cg12876356, cg13039251, cg15693572, cg06126421 e cg12803068 hanno mostrato il bias più elevato in assoluto. Complessivamente 8 CpG su 13 hanno presentato la tendenza a stimare erroneamente i livelli di metilazione previsti; inoltre, non è stata riscontrata una correlazione significativa tra l'errore di amplificazione rilevato e il numero di siti CpG contenuti nell'amplicone, la sua lunghezza e la profondità di sequenziamento. Sulla base dei dati ottenuti dagli standard, i bias sono quindi stati corretti tramite modelli statistici (bi-esponenziali) applicati per ogni CpG. La correzione ha mostrato un impatto negativo sull'accordo tra i duplicati tecnici e i campioni analizzati in precedenza; tuttavia, ciò è dovuto alla presenza di due outlier, senza i quali la correlazione migliora. In particolare,

cg12876356 ha mostrato l'impatto maggiore, con differenze tra le repliche fino al 62%. Dal momento che la correzione dei dati di metilazione potrebbe fornire risultati discordanti circa la predizione dello stato di fumatore, per completezza è stato deciso di eseguire l'analisi su entrambi i set di dati non corretti e corretti.

3.2 Assegnazione dei campioni in studio ai fenotipi del fumo utilizzando i modelli di previsione.

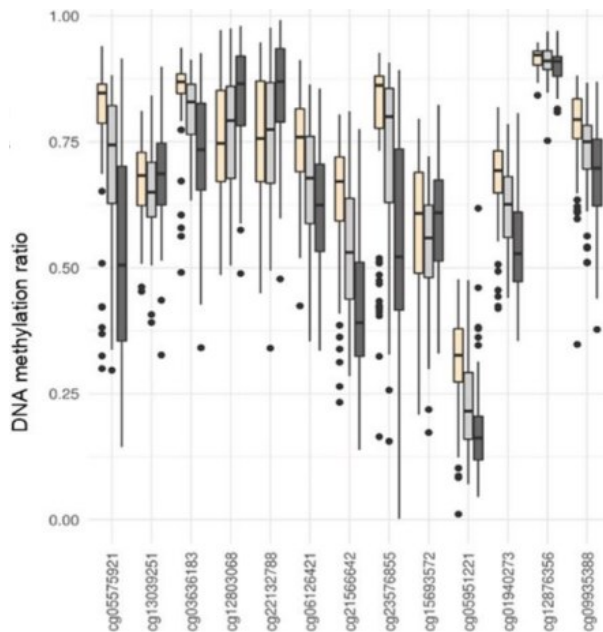


Figura 3: boxplot rappresentanti la distribuzione della metilazione per CpG tra le categorie di fumatori: i boxplot gialli rappresentano la distribuzione della metilazione nei mai fumatori, quelli grigi la distribuzione della metilazione nei fumatori attuali e quelli neri la distribuzione della metilazione nei mai fumatori.

Utilizzando il nuovo test MPS e basandosi sui valori di metilazione non corretti, la stragrande maggioranza dei campioni di fumatori attuali e mai fumatori sono stati raggruppati separatamente, anche se ciò risulta più complesso per gli ex fumatori. Sono state riscontrate prove della transizione della metilazione da fumatori attuali a ex fumatori a mai fumatori, per la maggior parte delle CpG non corrette dai bias di PCR e ciò risulta essere in accordo con l'analisi svolta da Maas et al. [5] basata su Microarray. Esaminando i valori di metilazione corretti dai bias di amplificazione, i risultati mostrano una tendenza analoga: la direzione e le firme

di metilazione risultano simili con cg12876356 che mostra il maggiore bias di amplificazione. In Figura 3 vediamo come per la maggior parte delle CpG, sono state osservate variazioni inter-individuali dei pattern di metilazione all'interno delle varie categorie: le fluttuazioni dinamiche dei livelli di metilazione sono significativamente maggiori nelle categorie di fumatori attuali ed ex fumatori, rispetto a quelle riscontrate nei mai fumatori; nonostante l'analisi sia stata svolta su 13 CpG considerate fortemente correlate al fumo, per cg13039251, cg12876356 e cg15693572, la distribuzione della metilazione risulta sovrapponibile tra le 3 categorie, motivo per cui, una volta effettuata l'analisi e ottenuti i dati sulla metilazione delle CpG, risulta difficile assegnare il campione con certezza ad una delle 3 categorie. Per cg12803068 e cg22132788, la distribuzione della metilazione risulta essere sovrapponibile per le categorie di mai fumatori ed ex fumatori, mentre è abbastanza diversificata per i fumatori attuali.

Per una futura analisi, sarebbe utile scegliere delle CpG la cui distribuzione della metilazione sia più diversificata per i 3 fenotipi del fumo.

3.3 Caratteristiche aggiuntive del pool di marcatori delle 13CpG.

Le variazioni inter-individuali dei pattern di metilazione all'interno delle varie categorie suggeriscono anche una correlazione tra i livelli di metilazione, l'intensità del fumo e la sua durata, motivo per cui si è cercato di testare questa correlazione. In particolare, lo studio si approfondisce tramite la *Figura 4* e la *Figura 5*.

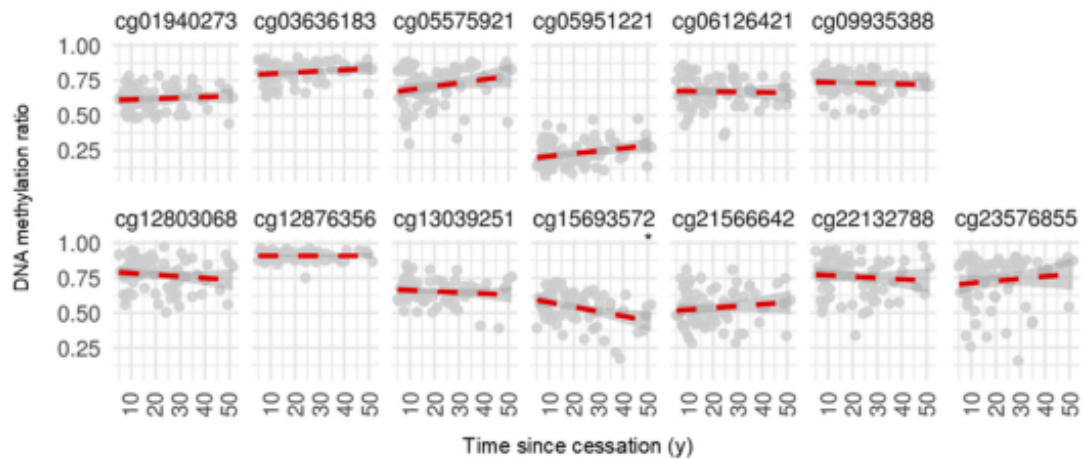


Figura 5: correlazione tra livello di metilazione delle CpG del fumo con il tempo trascorso dalla cessazione dal fumo negli ex fumatori, utilizzando dati ottenuti da MPS privi di correzione di bias

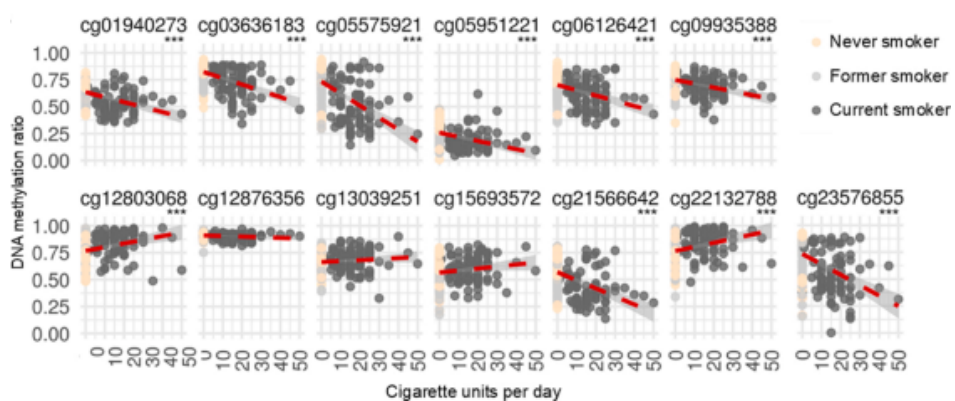


Figura 4: correlazione tra il livello di metilazione e le sigarette fumate al giorno nei fumatori attuali, utilizzando dati ottenuti da MPS privi di correzione dai bias

Riferendoci al grafico presente in *Figura 4*, si riscontra un'associazione statisticamente significativa, anche se debole, solo per cg15693572: dai dati ottenuti si evince che è possibile recuperare la corretta metilazione, in dipendenza del tempo trascorso dalla cessazione dal fumo. Il grafico presente in *Figura 5*, invece, dimostra che 10 su 13 CpG relative al fumo rilevano una forte correlazione statisticamente significativa con le sigarette fumate al giorno, ad eccezione di

cg12876356, cg13039251 e cg15693572, dai cui si evince che il cui cambiamento nel pattern di metilazione, nei fumatori attuali, è di tipo quantitativo.

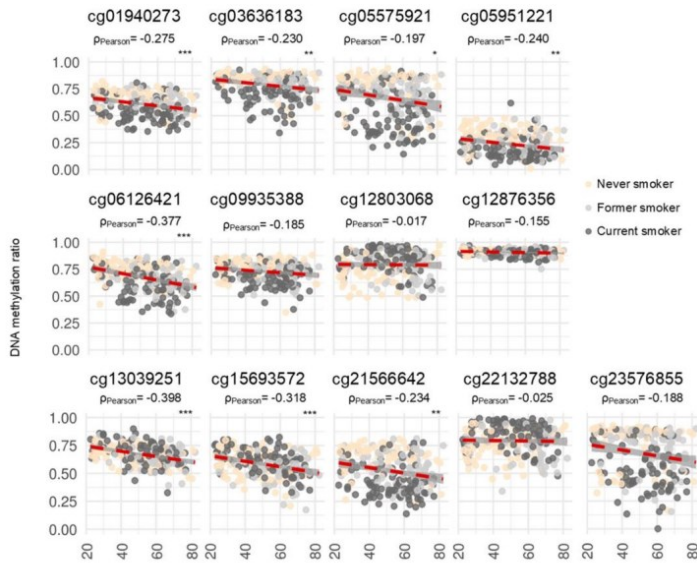


Figura 6: influenza dell'invecchiamento sulle CpG del fumo

La metilazione del DNA legata al fumo, è stata anche associata a un'accelerazione epigenetica dell'età, in particolare, in organi come i polmoni; questa accelerazione non sembra rallentare a seguito della cessazione dal fumo. Per questo motivo, sono stati analizzati i potenziali effetti aggiuntivi del fumo sull'invecchiamento (Figura 6).

Nei grafici in Figura 6, si nota un'associazione negativa statisticamente significativa tra metilazione delle CpG ed età cronologica. Ulteriori analisi sugli effetti epigenetici nel fumo a livello dei polmoni hanno dimostrato un effetto epigenetico del fumo sesso-specifico, le donne appaiono più suscettibili al fumo di sigaretta e spesso sviluppano malattie polmonari più gravi.

3.4 Previsione dell'abitudine al fumo utilizzando modelli statistici vecchi e nuovi.

Per fare previsioni sull'abitudine al fumo a partire dall'analisi delle CpG del sangue, usando MPS, sono stati utilizzati modelli statistici (modello di regressione logistica multivariata a 2 categorie e modello di regressione logistica multinomiale a 3 categorie) basati su Microarray pubblicati in precedenza da Maas et al. [5] e applicati ai dati MPS ottenuti. Successivamente sono stati misurati precisione e sensibilità del test. Testando i modelli sui dati MPS non corretti e corretti dai bias, si nota un calo di precisione della predizione nel momento in cui si testano i modelli utilizzando i dati MPS corretti dai bias. In generale, l'accuratezza è stata sostanzialmente ridotta utilizzando i dati MPS, a causa delle differenze di metilazione introdotte dalla tecnologia, ma anche dalla mancanza di validazione delle informazioni del questionario nel nostro studio. Inoltre, la dimensione del campione per i dati MPS è significativamente più piccola rispetto ai dati Microarray, ciò nonostante si è cercato di tenere conto dei bias tecnici nei parametri del modello. E' stato verificato in modo simulato come questi parametri cambiano rispetto ai dati di Microarray creando dei modelli congiunti (4 modelli) per la predizione a 2 e 3 categorie, utilizzando come strategia quella di costruire

modelli separatamente su ciascuna tecnologia e combinarli in un unico modello per confrontare i risultati ottenuti fianco a fianco. Si è notato come la predizione ottenuta utilizzando i dati Microarray non è stata influenzata dall'aggiunta dei dati MPS grezzi o corretti:

- Nel modello a 2 categorie sono stati assegnati correttamente il 97.6% dei mai fumatori e il 58.5% dei fumatori attuali
- Nel modello a 3 categorie le predizioni corrette corrispondevano al 78% sui mai fumatori, 65.2% sugli ex fumatori e 66.8% sugli attuali fumatori

Per il sottoinsieme di dati MPS non corretti sono stati ottenuti i seguenti risultati:

- Nel modello a 2 categorie sono stati assegnati correttamente l'86.6% dei mai fumatori e il 64.4% dei fumatori attuali
- Nel modello a 3 categorie le previsioni corrette corrispondono al 73.2 % dei mai fumatori, il 50.7% degli ex fumatori e il 71.1% dei fumatori attuali.

Nella determinazione della probabilità di assegnazione di ogni campione ad una categoria è stata osservata una mancanza di campioni tra fumatori attuali e mai fumatori i quali andavano ad essere assegnati agli ex fumatori, e questo errore di assegnazione si ha in entrambi i modelli. Elevata concordanza nei risultati di previsioni si ha anche con le repliche tecniche. Infine, una performance simile è stata ottenuta per i modelli MPS corretti: grazie al sottoinsieme di dati MPS, l'accuratezza di previsione ottenuta è stata notevolmente migliorata rispetto ai modelli basati solo su Microarray.

CAPITOLO 4: DISCUSSIONE

La profilazione della metilazione del DNA risulta essere un metodo promettente non solo per indagare l'abitudine al fumo, ma anche per indagare tratti dello stile di vita determinati dall'esposizione ambientale. In precedenza, sono stati fatti molti studi con lo scopo di creare modelli predittivi per l'abitudine al fumo:

- Shenker et al. [6] hanno utilizzato il pirosequenziamento bisolfito su 4 loci genomici che presentavano livelli di metilazione strettamente legati all'esposizione al tabacco. Poiché il pirosequenziamento è usato principalmente in forma single-plex e necessita di un numero di cicli di PCR elevato, risulta essere svantaggioso: per condurre l'analisi, sarebbe necessaria una traccia di DNA abbondante e di buona qualità; inoltre il numero enorme dei cicli di amplificazione, aumenterebbe i bias tecnici.
- Philibert et al. [7] hanno utilizzato la digital PCR per analizzare una singola CpG del fumo. Nonostante la digital PCR fornisca dati di metilazione molto accurati, questo approccio offre una bassa capacità multiplex.
- Kondratyev et al. [8] hanno usato sistemi di sequenziamento NGS a singola molecola (PacBio su prodotti di long PCR trattati con bisolfito) per lo studio di effetti di metilazione allele-specifici: sono stati usati cinque bersagli noti per essere influenzati dal fumo e dall'analisi di questi bersagli sono stati identificati significativi effetti di metilazione allele-specifici nelle

regioni *AHRR* e *IER3*: queste informazioni potrebbero essere sfruttate per migliorare la previsione del fumo sulla base dei dati di metilazione del DNA raccolti. Tuttavia il sequenziamento con PacBio risulta costoso e l'analisi di soli 5 target non determina reali vantaggi dal punto di vista forense.

- Più recentemente Wen et al. [9] hanno proposto un saggio basato sull'estensione di primer a singolo nucleotide sensibile alla metilazione per lo studio di 9 loci genomici legati al fumo. Questo approccio è stato usato per altre applicazioni forensi come determinare il tipo di tessuto di una traccia biologica umana, stimare l'età di un donatore di tracce sconosciuto e differenziare gemelli monozigoti. Il sistema proposto offre anche dei vantaggi metodologici rispetto ad MPS, in particolare, quando si analizza un numero ridotto di campioni.

4.1 Nuovo modello basato su MPS e i suoi limiti

Il sequenziamento MPS offre vantaggi convincenti rispetto alle tecniche di rilevamento standard, in particolare, per quanto riguarda sensibilità, capacità multiplex e risoluzione; nonostante ciò, sono stati riscontrati problemi durante la progettazione, amplificazione e sequenziamento della multiplex PCR.

La conversione con bisolfito non sembra essere adatta alla tipizzazione forense, in quanto necessita di grandi quantità di DNA genomico (dai 200 ai 500 ng) e può contribuire a degradare ulteriormente del DNA già degradato. Inoltre, a causa della sua frammentazione, dello stato a singolo filamento e con regioni a bassa complessità (stretch di T), il DNA convertito risulta essere difficile da amplificare: i problemi riscontrati riguardano la difficoltà nel disegnare primers specifici, che rende quindi difficile amplificare le 13CpG in un'unica reazione di multiplex PCR, con lo scopo di ridurre i costi e le tempistiche di analisi. Nonostante ciò, gli autori sono riusciti a riunire le 13 CpG in 3 multiplex PCR che, pur limitando il numero di reazioni, non rappresenta la condizione ottimale per una analisi di tipo forense. Per questi motivi appena elencati, un lavoro futuro dovrebbe proporre l'ottimizzazione della fase di conversione nonché di amplificazione per garantire maggiore sensibilità ed efficienza del sistema.

Un altro problema riscontrato durante l'analisi dei campioni riguarda la presenza di bias durante l'amplificazione, che, nonostante siano stati corretti tramite modelli statistici (bi-esponenziali), non hanno permesso di migliorare l'accuratezza di rilevamento della metilazione e l'accuratezza della previsione dell'abitudine al fumo. Il successo della strategia di correzione dell'amplificazione PCR dipende fortemente dal livello di metilazione osservato per ogni marcatore e dal contributo di ogni CpG al modello di previsione. Si può risolvere il problema del bias riprogettando i primers di PCR, ma questo risulterebbe essere molto impegnativo per frammenti molto densi di CpG in quanto si è costretti ad includere basi degenerate nella sequenza dei primers per evitare un legame preferenziale. In

alternativa, per limitare i bias di amplificazione, si potrebbe usare un ciclo di PCR ridotto (30-32 cicli) per ridurre al minimo gli effetti tra i vari saggi, a condizione che la quantità di materiale posta in input soddisfi quella richiesta per la preparazione della libreria. Un lavoro futuro dovrebbe esplorare delle strategie di correzione aggiuntive: un esempio sono gli identificatori molecolari unici (UMI) legati alle molecole di RNA nella prima fase della preparazione della libreria di sequenziamento. Questi hanno lo scopo di stabilire un'identità distinta per ogni molecola in input: dopo la PCR, si presume che molecole che condividono lo stesso UMI derivino dalla stessa molecola in input, in modo tale da riuscire a discriminare eventuali aspecifici che tendono a formarsi quando sono richiesti molti cicli di PCR. Per evitare ulteriori bias di amplificazione, nello studio analizzato in questo elaborato, è stata costruita una libreria genomica priva di amplificazione, utilizzando il massimo input possibile di prodotto di PCR (500ng). In questo modo è stato ottenuto un coverage di sequenziamento sufficiente e bilanciato tra i vari ampliconi, e ciò ha portato ad un rilevamento della metilazione riproducibile. Sono comunque stati rilevati potenziali bias post-PCR, quindi sono necessarie altre ricerche per determinare il coverage minimo di lettura per ogni CpG al fine di poter analizzare simultaneamente in un'unica corsa un maggior numero di campioni, riducendo così tempi e costi dell'analisi. È quindi fondamentale un'implementazione del saggio MPS per aumentare sensibilità e robustezza in modo tale da poter utilizzare questo metodo per le analisi forensi future.

4.2 Scelta delle CpG e limiti di classificazione

L'analisi di sole 13 CpG del fumo, selezionate dallo studio di Maas et al. [5], risulta essere un ulteriore limite riscontrato nello studio in analisi. In studi successivi potrebbe essere utile indagare altre CpG, in modo da sostituirle a quelle che sono risultate essere debolmente correlate con l'abitudine al fumo.

Ulteriore limite dello studio è quello di cercare di classificare i fumatori in 2 o 3 categorie distinte: il fumo risulta essere un tratto piuttosto quantitativo e ciò è stato dimostrato dallo studio stesso, il quale evidenzia un'associazione significativa tra il livello di metilazione delle CpG e il numero di sigarette fumate al giorno. Risulta essere meno significativa l'associazione tra il livello di metilazione delle CpG in analisi e il tempo trascorso dalla cessazione dal fumo, ma ciò probabilmente è dovuto alla ridotta sensibilità del sistema. Alcuni studi hanno indagato su quest'ultimo aspetto:

- McCartney et al. [10] hanno riscontrato che il tempo di esposizione al fumo, a partire dal quale almeno il 50% dei fumatori attuali veniva assegnato al cluster dei fumatori attuali, varia dai 5 ai 9 anni nei fumatori più assidui e dai 15 ai 19 anni nei fumatori più leggeri. Inoltre gli ex fumatori

a basso dosaggio, hanno una maggiore possibilità di rientrare nella categoria di non fumatori a partire da un anno dopo la cessazione dal fumo (a differenza degli ex fumatori ad alto dosaggio che necessitano di almeno 2 anni);

- Philibert et al. [11] hanno testato la possibilità che la metilazione di una CpG potesse essere usata per verificare la cessazione dal fumo: sono stati determinati i livelli mensili di metilazione di una CpG, in un gruppo di 67 fumatori auto-segnalati, sottoposti a terapia per smettere di fumare monitorata biochimicamente tramite cotinina, sostanza usata come biomarcatore per la quantificazione dell'esposizione al fumo attivo e passivo di tabacco, in quanto permane a lungo nell'organismo ed è possibile dosarla, oltre che nel sangue, anche nella saliva e nell'urina. È risultato che, in 20 soggetti su 60 che hanno completato il protocollo di tre mesi, la reversione della metilazione della CpG in esame dipende dall'intensità del fumo iniziale, con livelli di metilazione nei fumatori più accaniti che tornavano alla media di 0.12% al giorno durante il periodo di trattamento di 3 mesi.

Considerati questi risultati, in ambito forense è quindi più utile fare una predizione a 2 categorie (fumatori attuali e non) in quanto in questo modo ci si concentra sulle abitudini attuali; per fare ciò bisogna comprendere prima tutti i comportamenti legati al fumo e, quindi, identificare gli ex fumatori. Dallo studio condotto, infine è emerso come il fumo passivo, sui non fumatori, modifichi il pattern di metilazione delle CpG del fumo al punto tale che questi risultino rientrare, tramite l'analisi delle CpG e l'applicazione dei modelli predittivi, all'interno del cluster dei fumatori attuali.

4.3 Altri fattori che influiscono sull'analisi condotta

Nello studio esaminato in questo elaborato, ci si è concentrati su dati provenienti dagli europei in quanto sono state utilizzate le CpG associate al fumo precedentemente identificate in campioni di popolazione europea. Tuttavia, l'associazione tra fumo e metilazione del DNA delle cellule del sangue (anche la dipendenza dalla dose) può essere specifica per ogni popolazione: per lo stesso numero di sigarette fumate, i nativi hawaiani, rispetto ai bianchi, hanno maggiore rischio di sviluppo di cancro ai polmoni, e i giapponesi americani hanno minore rischio di sviluppo di cancro. L'influenza della dose di fumo sui modelli di metilazione del DNA risulta eterogenea tra razze/etnie. Studi successivi dovrebbero concentrarsi sulla valutazione di tali differenze e sull'identificazione di biomarcatori che catturino meglio gli effetti del fumo in modo trasversale nelle diverse popolazioni umane.

Un altro limite dello studio in esame è quello di essersi concentrati solamente sui livelli di metilazione del sangue; ciò è stato dovuto al fatto che i marcatori predittivi impiegati sono stati precedentemente identificati nel sangue. Poiché la metilazione associata al fumo può avere livelli diversi a seconda del tipo di cellule prese in considerazione, in futuro, sarebbe interessante studiare come la conoscenza dei livelli di metilazione legati al fumo di vari tipi cellulari possa influire sulla previsione dell'abitudine al fumo. In relazione a ciò sarebbe utile anche indagare i livelli di metilazione legati al fumo in tessuti di rilevanza forense quali ad esempio saliva, liquido seminale e sudore.

4.4 Pattern di metilazione e stile di vita

Un aspetto fondamentale rilevato dallo studio è l'associazione tra i livelli di metilazione delle CpG in esame e l'età: i risultati suggeriscono che l'esposizione al fumo attivo accelera l'età di metilazione del DNA nel sangue in maniera quantitativa e quindi associata a pacchetti di sigarette fumate all'anno; in particolare, i risultati ottenuti nello studio in esame, suggeriscono una riduzione della metilazione nelle CpG del fumo, che rimane anche dopo aver corretto l'abitudine al fumo. Questo risulta fondamentale nelle analisi forensi, in quanto l'analisi del pattern di metilazione può rilevare anche informazioni aggiuntive sull'età del donatore anche se, il valore predittivo dell'età a partire dall'analisi delle CpG del fumo, deve essere sottoposto a valutazione.

Ad influenzare i livelli di metilazione sono anche le condizioni ambientali in cui i soggetti in studio vivono: si è notato che la metilazione di una CpG è più bassa negli adulti non fumatori che però risiedono in aree ad alto tasso di inquinanti atmosferici.

Conoscendo la stretta correlazione tra il pattern di metilazione e lo stile di vita, per migliorare la comprensione di tratti complessi e garantire un'applicazione pratica dell'analisi di metilazione delle CpG, in ambito forense, sarebbe utile esplorare l'associazione del fumo con tratti come l'indice di massa corporea: sembra infatti che l'obesità abbia radici genetiche; inoltre i geni legati all'obesità, possono anche interagire con fattori ambientali. Avere la possibilità di predire in maniera corretta, a partire dal pattern di metilazione, informazioni come età e caratteristiche legate allo stile di vita permetterebbe di ottenere un'impronta epigenomica personalizzata e di migliorare l'accuratezza della previsione.

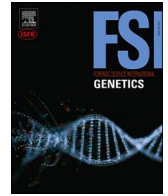
CONCLUSIONI

Lo studio esaminato in questo elaborato si propone lo scopo di sviluppare un nuovo saggio basato sul sequenziamento MPS per studiare il pattern di metilazione indotto dal fumo nel sangue. In particolare, sono state analizzate 13

CpG associate al fumo rilevate in precedenza da Maas et al. [5] La valutazione del metodo con duplicati tecnici e controlli metilati artificialmente ha rivelato che il saggio MPS sviluppato è accurato e riproducibile, ma soffre anche di un noto bias di amplificazione. Il nuovo metodo è stato applicato ad una serie relativamente ampia di campioni di sangue provenienti da donatori; sono state poi esaminate le differenze di metilazione tra fumatori abituali, ex fumatori e mai fumatori, associando i livelli di metilazione ai comportamenti legati al fumo, come le sigarette giornaliere e tempo trascorso dalla cessazione e ad altre caratteristiche come l'età e il sesso. È risultato che l'accuratezza della previsione, per entrambi i modelli a due e tre categorie, è stata mantenuta riqualificando i dati microarray precedenti con i dati MPS generati di recente, per tenere conto dei bias determinati dalla tecnologia e correggerli. In sintesi, il nuovo metodo sviluppato può essere utilizzato per studiare ulteriormente i modelli di metilazione associati al fumo nel sangue e in altri tessuti e questo permette di avvicinarsi alle future applicazioni forensi.

BIBLIOGRAFIA

- [1] L. P. Breitling, R. Yang, B. Korn, B. Burwinkel, and H. Brenner, "Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication," *The American Journal of Human Genetics*, vol. 88, no. 4, pp. 450–457, Apr. 2011, doi: 10.1016/j.ajhg.2011.03.003.
- [2] K. Sugden *et al.*, "Establishing a generalized polyepigenetic biomarker for tobacco smoking," *Transl Psychiatry*, vol. 9, no. 1, p. 92, 2019, doi: 10.1038/s41398-019-0430-9.
- [3] D. L. McCartney *et al.*, "Epigenetic prediction of complex traits and death," *Genome Biol*, vol. 19, no. 1, p. 136, 2018, doi: 10.1186/s13059-018-1514-1.
- [4] S. Bollepalli, T. Korhonen, J. Kaprio, S. Anders, and M. Ollikainen, "EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data," *Epigenomics*, vol. 11, no. 13, pp. 1469–1486, Aug. 2019, doi: 10.2217/epi-2019-0206.
- [5] S. C. E. Maas *et al.*, "Validated inference of smoking habits from blood with a finite DNA methylation marker set," *Eur J Epidemiol*, vol. 34, no. 11, pp. 1055–1074, 2019, doi: 10.1007/s10654-019-00555-w.
- [6] N. S. Shenker *et al.*, "DNA Methylation as a Long-term Biomarker of Exposure to Tobacco Smoke," *Epidemiology*, vol. 24, no. 5, 2013, [Online]. Available: https://journals.lww.com/epidem/fulltext/2013/09000/dna_methylation_as_a_long_term_biomarker_of.12.aspx
- [7] R. Philibert *et al.*, "Dose Response and Prediction Characteristics of a Methylation Sensitive Digital PCR Assay for Cigarette Consumption in Adults," *Front Genet*, vol. 9, 2018, doi: 10.3389/fgene.2018.00137.
- [8] N. Kondratyev, A. Golov, M. Alfimova, T. Lezheiko, and V. Golimbet, "Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation," *Clin Epigenetics*, vol. 10, no. 1, p. 130, 2018, doi: 10.1186/s13148-018-0565-1.
- [9] D. Wen *et al.*, "DNA methylation analysis for smoking status prediction in the Chinese population based on the methylation-sensitive single-nucleotide primer extension method," *Forensic Sci Int*, vol. 339, p. 111412, 2022, doi: <https://doi.org/10.1016/j.forsciint.2022.111412>.
- [10] D. L. McCartney *et al.*, "Epigenetic signatures of starting and stopping smoking," *EBioMedicine*, vol. 37, pp. 214–220, Nov. 2018, doi: 10.1016/j.ebiom.2018.10.051.
- [11] R. Philibert *et al.*, "The Reversion of cg05575921 Methylation in Smoking Cessation: A Potential Tool for Incentivizing Healthy Aging," *Genes (Basel)*, vol. 11, no. 12, 2020, doi: 10.3390/genes11121415.



Research paper

Targeted DNA methylation analysis and prediction of smoking habits in blood based on massively parallel sequencing

Athina Vidaki^{a,*}, Benjamin Planterose Jiménez^a, Brando Poggiali^a, Vivian Kalamara^{a,1}, Kristiaan J. van der Gaag^b, Silvana C.E. Maas^{a,c,2}, B.I.O.S. Consortium, Mohsen Ghanbari^c, Titia Sijen^{b,d}, Manfred Kayser^a

^a Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

^b Division of Biological Traces, Netherlands Forensic Institute, The Hague, the Netherlands

^c Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

^d Swammerdam Institute of Life Sciences, University of Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:

Forensic epigenetics
DNA methylation
Smoking
Lifestyle prediction
Massively parallel sequencing
Blood

ABSTRACT

Tobacco smoking is a frequent habit sustained by > 1.3 billion people in 2020 and the leading preventable factor for health risk and premature mortality worldwide. In the forensic context, predicting smoking habits from biological samples may allow broadening DNA phenotyping. In this study, we aimed to implement previously published smoking habit classification models based on blood DNA methylation at 13 CpGs. First, we developed a matching lab tool based on bisulfite conversion and multiplex PCR followed by amplification-free library preparation and targeted paired-end massively parallel sequencing (MPS). Analysis of six technical duplicates revealed high reproducibility of methylation measurements (Pearson correlation of 0.983). Artificially methylated standards uncovered marker-specific amplification bias, which we corrected via bi-exponential models. We then applied our MPS tool to 232 blood samples from Europeans of a wide age range, of which 90 were current, 71 former and 71 never smokers. On average, we obtained 189,000 reads/sample and 15,000 reads/CpG, without marker drop-out. Methylation distributions per smoking category roughly corresponded to previous microarray analysis, showcasing large inter-individual variation but with technology-driven bias. Methylation at 11 out of 13 smoking-CpGs correlated with daily cigarettes in current smokers, while solely one was weakly correlated with time since cessation in former smokers. Interestingly, eight smoking-CpGs correlated with age, and one displayed weak but significant sex-associated methylation differences. Using bias-uncorrected MPS data, smoking habits were relatively accurately predicted using both two- (current/non-current) and three- (never/former/current) category model, but bias correction resulted in worse prediction performance for both models. Finally, to account for technology-driven variation, we built new, joint models with inter-technology corrections, which resulted in improved prediction results for both models, with or without PCR bias correction (e.g. MPS cross-validation F₁-score > 0.8; 2-categories). Overall, our novel assay takes us one step closer towards the forensic application of viable smoking habit prediction from blood traces. However, future research is needed towards forensically validating the assay, especially in terms of sensitivity. We also need to further shed light on the employed biomarkers, particularly on the mechanistics, tissue specificity and putative confounders of smoking epigenetic signatures.

1. Introduction

Human genetic variation, for example via short tandem repeats (STRs), allows to identify individuals with very high discriminatory

power [1]. This is only possible when the genetic profile at the crime scene matches the one from a law enforcement database or a reference provided in the case. Whenever this is not the case, due to the comparative nature of standard STR profiling, the investigation is led

* Corresponding author.

E-mail address: a.vidaki@erasmusmc.nl (A. Vidaki).

¹ Present address: Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Heraklion, Greece.

² Present address: Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.

<https://doi.org/10.1016/j.fsigen.2023.102878>

Received 16 December 2022; Received in revised form 28 March 2023; Accepted 18 April 2023

Available online 20 April 2023

1872-4973/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

into a deadlock. Consequently, additional genetic variation, for example via single nucleotide polymorphisms (SNPs), may allow for the prediction of biogeographical ancestry and appearance traits (forensic DNA phenotyping, FDP) [2]; hence, providing the police with investigative leads that can narrow down the pool of suspects. Until now, the most accurate and widely used FDP markers are SNPs predictive for eye, hair, skin color, and biogeographic ancestry [3], while the first models and tools are already available for additional appearance traits such as freckles, eyebrow color, hair loss in men and hair structure [4].

In contrast to genetic variation, epigenetics has only recently been proposed [5] and is being increasingly explored in forensics [6]. More specifically, DNA methylation is the most studied epigenetic biomarker, involving the addition of a methyl-group to carbon-5 of cytosine nucleotides followed by guanines (CpGs). While the vast majority of CpGs are methylated in most human cells, methylation at gene promoters can influence gene expression, in a quantitative and dynamic way [7]. Hence, forensic epigenetic profiling offers great avenues to study not only distinct cell type-specific processes, but also (dynamic) environmentally influenced phenotypic traits [6]. For example, methylation profiling is promising for the confirmatory identification of body fluids and tissues [8,9], prediction of chronological age [10–12] and discrimination of identical (monozygotic) twins [13,14]. So far, a limited number of forensically relevant CpG markers have been discovered and applied to such forensic epigenetic applications, mainly using technologies that enable the analysis of a small number of CpG markers [6]. Nevertheless, it was recently proposed that future forensic epigenomic profiling may allow the prediction of additional traits, such as lifestyle habits [15]. Such prediction will potentially lead to a broadened FDP but will also allow us to achieve more accurate age prediction, as lifestyle factors are known to confound ageing signatures used for chronological age prediction [16].

Particularly, tobacco smoking is a widely frequent lifestyle habit, with 1.3 billion smokers worldwide in 2020 [17]; despite being widely established as a major health risk [18], even a long time after cessation [19]. Overall, this corresponds to 22.3 % of the global adult population (36.7 % in men and 7.8 % in women), with > 80 % living in low- and middle-income countries [17]. Specifically in Europe, 5–30 % of the population frequently sustains smoking habits, depending on the country [20]. As a result, smoking is one of the leading disease risk factors and is associated with millions of deaths each year worldwide, both from direct and passive exposure [17]. Given its prevalence, predicting a person's smoking habits is relevant in public health and personalized medicine fields [21], particularly to validate electronic medical records or research participant questionnaires, which often contain incorrect or missing data [22,23]. Particularly in the forensic context, it may be useful to predict a current, frequent smoker, as such habit is often known and visible to individuals within one's surroundings. While toxicological analysis of nicotine and its metabolites (e.g. cotinine) can offer a solution, the specificity is low (high rate of false negatives) since such metabolites tend to have short half-lives (~15–19 h for cotinine); therefore, assessing current and acute, rather than habitual, smoking [24]. In addition to its high costs, their reduced scalability made researchers turn towards more promising biological approaches.

Smoking is known to induce not only hypoxia [25], DNA damage [26] and telomere shortening [27], but also substantial genome-wide epigenetic alterations, for example by affecting DNA methyltransferase expression [28]. Over the last decade, there are several epigenome-wide association studies (EWAS) suggesting smoking-associated methylation changes, mainly in blood [29–33]. Such studies have uncovered thousands of smoking-CpGs in several genes such as the aryl hydrocarbon receptor repressor (AHRR), F2R like thrombin or trypsin receptor 3 (F2RL3), alkaline phosphatase, placental-like 2 (ALPP2), growth factor independent 1 transcriptional repressor (GFII1), G protein-coupled receptor 15 (GPR15) and myosin 1 G (MYO1G). Smoking-induced single-CpG methylation changes are robust but relatively small (< 20 %),

and mainly result in decreased methylation levels in smokers [30]. They are associated with accumulative smoke exposure in current smokers, but they are also reversibly associated with time since cessation in former smokers [34,35], both in quantitative manner.

Genome-wide DNA methylation profiling, for example via the Illumina microarray platforms, allows researchers to discover phenotype-related blood methylation changes in population cohorts, and develop robust molecular predictors to be applied in practice to replace self-reported phenotypes [36]. Sugden et al. computed a standardized smoking score based on 2623 CpGs and validated it in two cohorts ($n_1 = 1037$; $n_2 = 2232$, respectively), which successfully discriminated never from both current/former smokers (area under the receiver operating characteristic (ROC) curve (AUC) range from 0.77 to 0.93) [37]. Moreover, McCartney et al. proposed an age- and sex-adjusted penalized regression model based on 233 smoking-CpGs, which explained a high proportion (> 60 %) of the variance in a large population cohort and discriminated the extremes: current ($n = 102$) from never ($n = 418$) smokers, with great accuracy (AUC = 0.98) [38]. Using a similar methodological approach Bollepalli et al. built a 121 CpG-based model for the prediction of current, former and never smokers and tested it in three independent datasets ($n_1 = 408$; $n_2 = 687$; $n_3 = 464$, respectively) [39]. Overall, current and never smokers were identified with 81 % sensitivity / 85 % specificity and 94 % sensitivity / 57 % specificity, respectively. For former smokers, their model showed a low average sensitivity of only 18 % but 96 % specificity [39]. Finally, with an attempt to reduce the smoking CpG marker set to a finite number, Maas et al. used data from six population cohorts ($n = 3764$) and built two- and three-category predictors based on only 13 smoking-CpGs, widely reported in the literature [40]. External model validation in an independent cohort ($n = 1608$) achieved an AUC of 0.91 for smokers/non-smokers and AUCs of 0.91 / 0.70 / 0.78 for current, former, never smokers, respectively. The Maas markers could also be used to successfully predict pack-years (number of packs of cigarettes smoked per day multiplied by the number of years the person has smoked) in current smokers (AUC = 0.80, 10 pack-years; AUC = 0.75, 15 pack-years) and smoking cessation time in former smokers (AUC = 0.76, 5 years; AUC = 0.76, 10 years; AUC = 0.75, 15 years).

Despite the promise of these smoking prediction models, it is challenging to employ the microarray technology on forensic-type samples, which are often of low DNA quantity / quality. Nevertheless, it is possible to develop a targeted PCR-based method when the number of smoking-CpGs is sufficiently small. Hence, the aim of this study was to develop a matching lab tool for the Maas model, benefiting at the same time from the promising massively parallel sequencing (MPS) technology in terms of marker multiplexing, resolution and sample throughput, which becomes increasingly more forensically relevant for a future application. To achieve our aim, 1) we developed a robust method based on bisulfite conversion and multiplex PCR followed by PCR-free library preparation and paired-end MPS on the Illumina MiSeq; 2) we assessed its performance using technical duplicates and artificially methylated standards to understand technical error and amplification bias, respectively; 3) we analyzed a relatively large set of whole blood samples ($n = 232$) from Europeans of a wide age range, including current, former and never smokers; 4) we correlated the detected methylation signatures with smoking habits and smoking-related traits, as well as other phenotypes, such as age and sex; 5) we tested both published Maas models for two- and three-category smoking habits prediction; and finally, 6) we built new corresponding, joint microarray:MPS models with technological corrections to account for inter-technological bias. To the best of our knowledge, this is the first study employing MPS technology for the prediction of smoking habits.

2. Materials & methods

The overall workflow of our study design is presented schematically in Fig. 1.

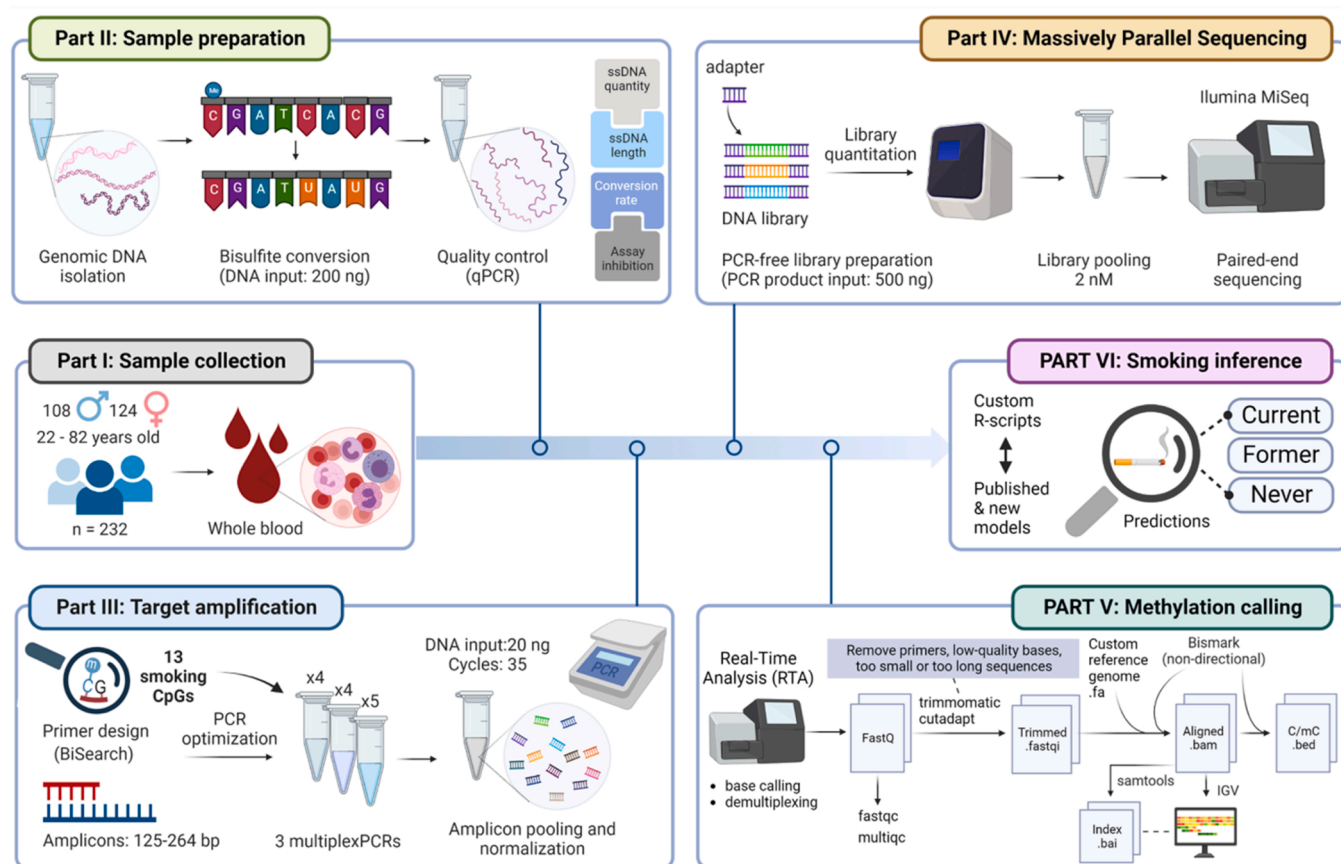


Fig. 1. Overview of key steps in our methodology from sample collection to smoking prediction.

2.1. Samples and controls

European individuals residing in the Netherlands were recruited as part of the family-based Erasmus Rucphen Family (ERF) study, following approval by the Medical Research Ethics (METC) Committee of Erasmus MC (213.575/2002/114). All donors provided written informed consent according to the Declaration of Helsinki. In our study, we included 232 participants, comprised of 108 males and 124 females aged from 22 to 82 years, grouped based on their smoking habits into 90 current, 71 former and 71 never smokers (Table 1). Smoking phenotypes (smoking habits, cigarettes smoked per day and age at smoking cessation) were collected using self-reported questionnaires. Table S1 contains detailed phenotypic information per participant. Peripheral venous blood was collected in Paxgene® Blood DNA tubes (QIAGEN, Hilden, Germany). Additionally, we prepared nine artificially methylated controls of different ratios (0, 10, 20, 40, 50, 60, 80, 90, 100 %) by mixing appropriate volumes from the human methylated and non-methylated DNA control set (Zymo Research, Irvine, California, USA).

Table 1
Summary of participants' phenotypic characteristics per smoking habits.

	Current	Former	Never	All
N	90	71	71	232
Sex (% female)	48.9	45.1	67.6	53.4
Age (y, mean ± SD)	52.8 ± 14.5	57.7 ± 14.3	46.0 ± 16.5	52.2 ± 15.7
Cigarettes per day (mean ± SD, N)	16.9 ± 9.1 (229)	0	0	NA
Cessation time (y, mean ± SD, N)	NA	21.1 ± 12.3 (70)	NA	NA

2.2. DNA sample preparation

Genomic DNA extraction was performed using the Paxgene® Blood DNA kit (QIAGEN) according to the manufacturer's instructions. Genomic DNA extracts were quantified in duplicate using the Quant-iT™ PicoGreen™ dsDNA Assay kit and Varioskan LUX Multimode Microplate Reader (ThermoFisher Scientific, Waltham, Massachusetts, USA) according to the manufacturer's recommendation. Average measurements were used to normalize DNA concentration to 4 ng/μL. Subsequently, a total of 200 ng (50 μL) of genomic DNA per sample was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research) according to the manufacturer's instructions following a 16-hour incubation. Bisulfite-converted DNA was eluted in 16 μL, while resulting concentrations and conversion efficiencies were indicatively evaluated in duplicate using a patent-pending qPCR assay [41]. Average measurements were used to normalize bisulfite DNA concentration to 10 ng/μL.

2.3. Smoking-associated CpGs

To predict smoking habits, we used the two- and three-categorical logistic and multinomial regression models from Maas et al., that employ the methylation on 13 smoking-CpGs to predict current versus non-current smokers and current versus former versus never smokers, respectively [40]. Detailed information on each smoking CpG's chromosomal location, DNA strand, closest gene and effects size (standardized odds-ratio on univariate models) are presented in Table S2. Microarray annotations were obtained from the IlluminaHumanMethylation450kanno. ilmn12.hg19 and Utools R-packages [42], while probe GRCh38/hg38 coordinates were obtained by transforming GRCh37/hg19 coordinates with the UCSC lift-over tool [43].

2.4. Bisulfite PCR

Ensembl genome browser (GRCh37/hg19) [44] was employed to extract the surrounding genomic sequences of the 13 smoking-CpGs. Bisulfite-converted genomic sequences were subsequently extracted using the web-based MethPrimer tool [45]. Bisulfite-specific primers were designed using the web-based BiSearch tool [46] using standard parameters. We particularly paid attention to design PCR assays that are shorter than 300 bp and as specific as possible, to account for the relatively fragmented and less complex nature of converted DNA template, respectively. Selected primer sets were scanned for potential dimers and hairpins using AutoDimer [47] and ordered from Integrated DNA Technologies (Coralville, Iowa, United States). Detailed information of our designed bisulfite PCR assays, including primer sequences, amplicon length and number of contained CpGs/total cytosines, are included in Table S2.

Single-plex bisulfite PCR assays were first optimized using annealing temperature, primer and $MgCl_2$ concentration gradients, and assessed using agarose gel electrophoresis. Subsequently, single-plex PCR reactions were combined based on shared annealing temperature; resulting multiplexes were further optimized by adjusting primer concentration to result in approximately equal amplification of all CpG markers. As a result, the 13 smoking-CpGs were split into three separate multiplex PCR reactions: PCR 1 with 4 CpGs (cg21566642, cg12876356, cg03636183 and cg01940273), PCR 2 with another 4 CpGs (cg05575921, cg23576855, cg15693572 and cg13039251) and PCR 3 with the final 5 CpGs (cg22132788, cg05951221, cg06126421, cg12803068 and cg09935388). Each multiplex PCR reaction was performed in a final volume of 20 μ L, containing 10 μ L of ZymoTaq™ PreMix (Zymo Research), 2 μ L of forward/reverse primer mix with assay-specific concentration as indicated in Tables S2, 1 μ L of $MgCl_2$ (25 nM), 6 μ L of H_2O and finally, 1 μ L of bisulfite-converted DNA (10 ng). The employed PCR cycling conditions were: 95 °C for 10 min; followed by 35 cycles of 95 °C for 30 s; 55 °C (PCR 1 and 2) or 54 °C (PCR 3) for 40 s; 72 °C for 1 min; and a final extension of 7 min at 72 °C. Successful PCR amplification was confirmed via agarose gel electrophoresis. Then, PCR products from all three multiplex PCR reactions were pooled together (final volume of 60 μ L) and purified using Agencourt AMPure XP beads (Beckman Coulter, Danvers, Massachusetts, United States) using 1 / 1.8 ratio in a final volume of 30 μ L of Resuspension buffer (Illumina, San Diego, California, United States). Purified PCR products were quantified in duplicate using the Quant-iT™ PicoGreen™ dsDNA Assay kit (ThermoFisher Scientific) according to the manufacturer's instructions. Average measurements were used to normalize DNA concentration to 10 ng/ μ L.

2.5. PCR-free library preparation and quantification

A total of 500 ng (50 μ L) of pooled and purified PCR amplicons per sample/control were used for library preparation in three separate experiments, that combined 78, 66 and 94 samples, respectively (summing up to 232 samples when excluding one sample from each of the six technical duplicate pairs). We opted for the TruSeq DNA PCR-Free High-Throughput Library Prep kit (96 samples) (Illumina) which includes no library amplification step, as an effort to reduce PCR amplification bias. Given that this kit is originally designed for epigenome-wide experiments, we did not perform the DNA fragmentation step since we were analyzing PCR amplicons of appropriate size. Other than that, the end-repair and library size selection, end adenylation and adapter ligation steps were performed according to the manufacturer's protocol. We used the Illumina unique dual indexes (IDT)– TruSeq DNA UD Indexes (96 Indexes, 96 Samples) (Illumina) during adapter ligation and Agencourt AMPure XP beads (Beckman Coulter) during clean-up steps. For the final purification, we used 20 μ L of Resuspension buffer (Illumina). Purified libraries were diluted in 1–5000 and 1–10,000 in Tris-HCl 10 mM / pH 8.5 with 0.1 % Tween-20 and quantified using the KAPA Library

Quantification Kit - Complete Universal for Illumina Platforms (Roche, Basel, Switzerland) following the manufacturer's instructions. Finally, indexed libraries were pooled together for a final concentration of 2 nM in a 300 μ L volume and the pool was quantified by an additional round of qPCR for determining the volume for optimal loading on the sequencer.

2.6. Targeted MPS based on Illumina MiSeq®

Using freshly made 0.2 N NaOH, pooled libraries were denatured and, through dilution with pre-chilled Hybridisation buffer (Illumina), 8–9 pM libraries were obtained. Finally, 20 % diluted PhiX control was added to the library and paired-end sequencing was performed using the 2 × 300-cycle MiSeq® reagent v3 cartridge (Illumina) on a MiSeq® FGx™ Forensic Genomics system (Illumina). Flow cell preparation and instrument set-up were performed per manufacturer's instructions.

2.7. MPS data analysis

MPS data was analyzed using our *in-house* bioinformatics pipeline Genomic Analysis by MPS on Bisulfite-converted Amplicons (GAMBA) (<https://github.com/BenjaminPlanterose/GAMBA>). GAMBA's input is a set of already de-multiplexed paired-end read files (.fastq), sequences for the set of amplicons (.fa) and primers (.fa). For all samples, GAMBA trims reads by primer sequences with trimmomatic [48] and filters by size (in this case, set to $100 \leq \text{size} \leq 230$) and quality score (default: Q₂₀) with cutadapt [49]. GAMBA does not perform de-duplication since this is an expected outcome from the amplicon nature of the experiment; thus, we assume that PCR primers do not contain unique molecular identifiers (UMIs). GAMBA calls fastQC [50] before and after read processing as part of the read quality control (QC). To summarize QCs across samples, GAMBA also calls multiqc [51]. For all pairs of fastq files, GAMBA performs bisulfite-converted read alignment and methylation calling with Bismark [52] on the non-directional library mode. This is required as we expect the complementary to the original top (CTOT) or bottom strand (CTOB), depending on whether a given primer was designed to target the original top (OT) or bottom (OB) strand, respectively (Table S2), unlike a typical whole-genome bisulfite sequencing directional library preparation protocol that specifically targets OT and OB. GAMBA combines methylated and unmethylated cytosine read counts across samples with custom R-scripts. Additionally, it sorts and indexes alignment files (.bam) with samtools [53], ready for input in the Interactive Genomics Viewer (IGV) [54].

2.8. Methylation analysis and prediction modelling

General data processing and visualization was performed in the R-programming language (version 4.2.1, 2022-06-23) [55] with R-packages data.table, ggplot2, gplots, RColorBrewer, scales, ggtern, ggribbles and corrplot. Model building and performance evaluation was carried out with R-packages stats, MASS, caret, nnet, minpack.lm, MLmetrics, ROCR, HandTill2001 and effectsize. To correct commonly encountered bisulfite PCR amplification bias as a result of different amplification efficiencies between methylated and non-methylated fragments that can heavily impact methylation quantification, we first fitted a bi-exponential model for each CpG (Table S3) with the Levenberg-Marquardt algorithm via minpack.lm::nlsLM based on calibration curves obtained from artificially methylated DNA standards. We then estimated PCR bias-free data by setting up an inverse problem and solving it with one dimensional root finding routine for each CpG and each individual (via stats::uniroot). Statistical inference on the dependence of CpG methylation with cigarettes per day, time since cessation, chronological age or sex was based on standard linear regression, correcting for covariates when needed. We corrected for three-category smoking status when testing for age and sex association, but not for time since cessation (only former smokers) and cigarettes per day

(phenotype of interest). We selected for significance at Bonferroni significant threshold to account for multiple testing correction. Two prediction models were made available by Maas et al.: a two-category multivariate logistic regression model (current versus non-current smoker), built with stats::glm with a “binomial” family and “log” link function and a three-category multivariate multinomial logistic regression model (current versus former versus never smoker), built with nnet::multinom (based on a softmax link function) [40]. We additionally built models for joint microarray:MPS data for two- and three-categories by employing the same fitting methods as Maas et al. To do so, we first pooled data from both technologies but adding an indicator covariate (either microarray or MPS); both with and without bias correction for the MPS data. In the parameter estimation, we allowed for technology-specific corrections in the intercept and the linear coefficients for all 13 CpGs to obtain models with a total of 28

(2-categories, Table S4) or 56 parameters (3-categories, Table S5), respectively. To test performance in an unbiased way, we implemented a repeated cross-validation strategy (5-fold, 2 repetitions) based on caret::train. We summarized 2×2 and 3×3 confusion matrices with the F₁-score or the weighted macro F-score, respectively, averaged across folds and repetitions when required. We have made available all employed scripts (https://github.com/BenjaminPlanterose/GAMBA/tree/main/others/scripts_employed_for_data_analysis) and models (<https://github.com/BenjaminPlanterose/GAMBA/tree/main/others/models>).

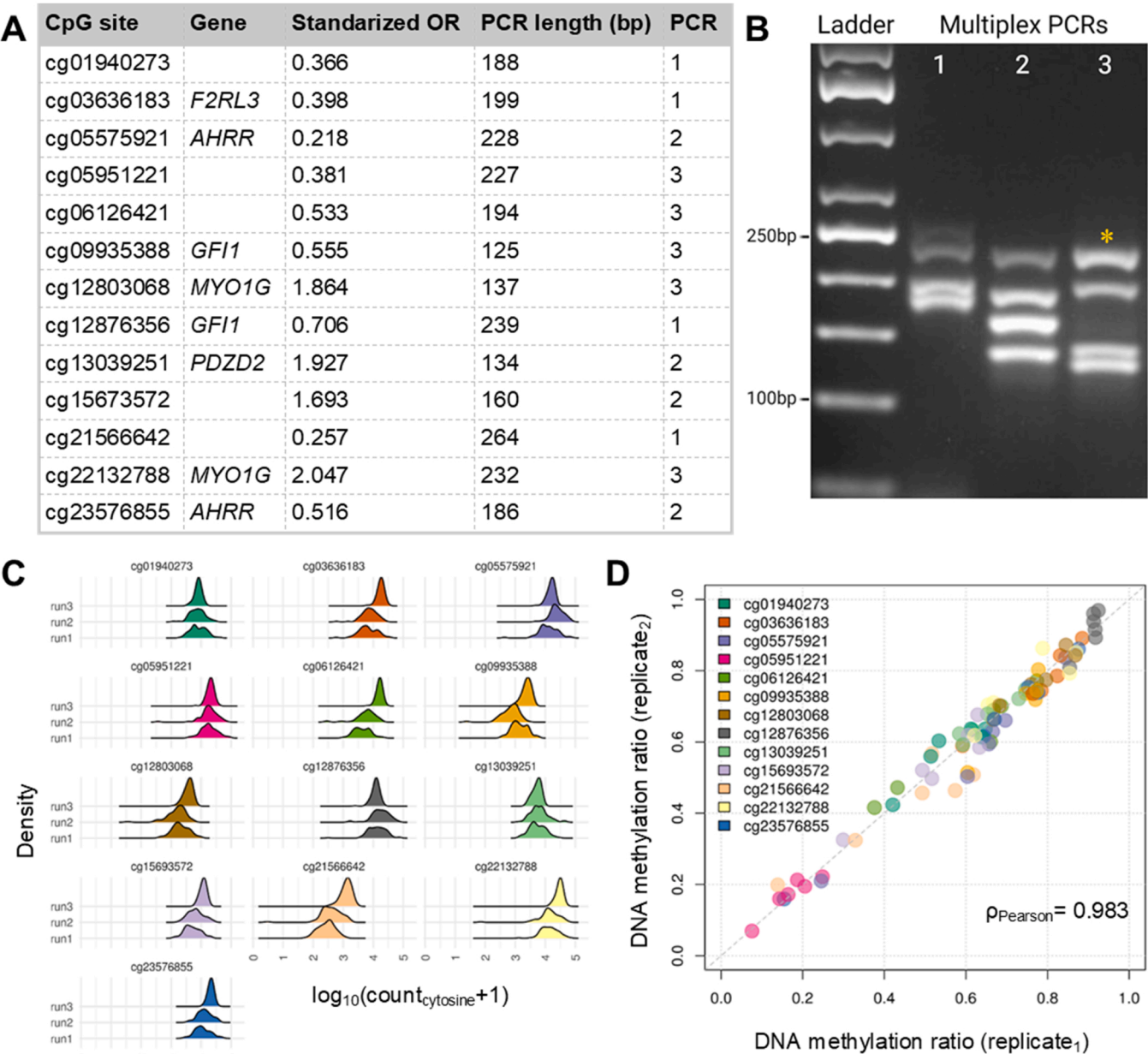


Fig. 2. Novel targeted methylation MPS assay. (A) Summary information on the employed 13 smoking-CpG markers; (B) Example agarose gel electrophoresis of the PCR products resulted from the three multiple bisulfite PCRs; (C) Ridge density plots showing read coverage per CpG / run; (D) Agreement in detected DNA methylation (without PCR bias correction) for six technical duplicates across CpGs. OR: Odds Ratio; ρ_{Pearson} : Pearson correlation coefficient; *: double PCR band corresponding to two assays (cg05951221 and cg22132788).

3. Results

3.1. Development of a novel targeted MPS assay to investigate smoking habits

To investigate smoking habits in blood in a targeted way, we aimed to develop a novel lab method to analyze 13 smoking-associated CpGs included in the microarray-based Maas models [40]. Our assay is based on bisulfite conversion and multiplex PCR followed by PCR-free library preparation and targeted paired-end MPS on the Illumina MiSeq® platform (Fig. 1). First, single-plex bisulfite PCR assays were designed and successfully optimized (amplicons of 125–265 bp, Fig. 2 A). Initially, we attempted to create a 13plex to amplify all CpGs and allow for more sensitive, cost-efficient and scalable analysis. However, this turned out challenging due to primer-dimer formation, non-specific or low amplification for certain bisulfite PCR assays. Hence, to account for the unequal performance among assays and to avoid missing data that could jeopardize our ability for predictions, we employed a diverse range of assay-specific primer concentrations (0.25–2.5 μ M) and instead formed three independent multiplex reactions (Fig. 2B, Table S2).

Generally, in an attempt to produce robust and reproducible methylation data for all samples included in our study, we employed optimal normalized conditions throughout our experimental pipeline, including for bisulfite conversion (genomic DNA input of 200 ng), bisulfite PCR (converted DNA input of 10 ng) and library preparation (PCR product input of 500 ng). Using this approach and our QC pipeline, we succeeded in obtaining good-quality, comparable performance metrics for all three MPS runs (Figs. S1–S3). More specifically, the total reads per MPS run were ≥ 30 million, resulting in an average of 188,828 read pairs per sample and 14,525 read pairs per CpG (Table S6), which exceeds the 1000 reads threshold for accurate methylation calling suggested in the literature [56]. Nevertheless, the average reads among CpGs varied significantly (range of 740 – 21,756 reads) across MPS runs (Fig. 2 C); with cg05951221 and cg22132788 being the highest performing, while cg09935388 and cg21566642 the worst performing CpG assays. Moreover, bisulfite-converted amplicons contained on average 50 non-CpG cytosines, which allowed us to calculate an overall bisulfite conversion efficiency on the amplified fragments ($>99\%$), indicating selective amplification of converted DNA strands (Figs. S1–S3).

Finally, we aimed to assess the reproducibility of methylation quantification based on our novel MPS assay by employing six sample duplicates that were randomly chosen from our blood DNA sample set. These technical duplicates were analyzed separately throughout the entire experimental pipeline (from the bisulfite conversion step onwards) and on different MPS runs. Overall, taking into account all CpGs and samples, the agreement in detected DNA methylation between technical duplicates was very high ($p_{\text{Pearson}} = 0.983$, Fig. 2D), with the average (\pm SD) absolute difference being only $3\% \pm 2.1\%$ (Table S7). Particularly, cg13039251 and cg23576855 displayed the lowest (both $1.5\% \pm 1.5\%$), while cg21566642 and cg05575921 the highest ($6.2\% \pm 4.2\%$ and $4.6\% \pm 3.2\%$, respectively) average (\pm SD) absolute difference. Interestingly, the cg21566642 assay also resulted in the lowest number of reads, which could at least be one of the reasons for the observed lower performance.

3.2. Evaluation of target amplification bias using artificially methylated DNA controls

Bisulfite PCR assays are known to suffer from different amplification rates between methylated and non-methylated amplicons, driven by the extensive sequence differences caused by the differential conversion of CpG sites. It is well known that PCR bias can potentially have a negative impact on the methylation quantification accuracy in downstream analysis [57]. To assess potential PCR amplification bias of our designed assays, we employed a set of artificially methylated DNA controls (0, 10, 20, 40, 50, 60, 80, 90, 100 %) that were analyzed in duplicate in two

different MPS experiments; except the 50 % standard that failed in the second run. For instance, Fig. S4 shows a read alignment example of cg05575921 for the 0 %, 50 % and 100 % artificially methylated standards. First, taking into account all CpGs, the inter-run agreement in detected DNA methylation across standards was very high ($p_{\text{Pearson}} = 0.986$), with the average (\pm SD) absolute methylation difference being only $3.3\% \pm 3.2\%$ (Table S8). While technical variation was similar to the one observed previously in the sample duplicates, this time cg22132788 and cg05951221 displayed the highest average (\pm SD) absolute difference ($11.7\% \pm 10.4\%$ and $6.3\% \pm 5.7\%$, respectively). Looking closely into the data per standard, it seems that this was driven by a larger detected methylation differences in intermediately-methylated samples (i.e. 20, 40, 60, 80 %), which are expected to suffer the most from minor micro-pipetting precision errors but mainly due to systematic amplification bias. On the methylation level, plotting the observed versus expected methylation at all CpG targets yielded deviation from linear detection for several assays, particularly cg12876356, cg13039251, cg15693572, cg06126421 and cg12803068 with the highest absolute bias (detected minus expected methylation summed across all samples: -3.79 to 4.8) (Fig. 3).

Overall, most assays (eight out of 13) had the tendency to underestimate expected methylation levels, but this was not correlated with PCR mixes. There was also no significant correlation between the detected amplification bias and the number of CpG sites contained in the amplicon (Pearson correlation test, p -value = 0.321), amplicon length (Pearson correlation test, p -value = 0.587) or average amplicon read depth (Pearson correlation test, p -value = 0.879). To account for PCR amplification bias we fitted bi-exponential models for each CpG based on the artificially methylated DNA standard data (Fig. S5A) and subsequently corrected our entire dataset with these models (Fig. S5B). PCR bias correction had a negative impact on the agreement between technical duplicates ($p_{\text{Pearson}} = 0.888$, Fig. S6C); however, this was driven by two outliers without which the correlation improves ($p_{\text{Pearson}} = 0.98$). The average (\pm SD) absolute difference was $4.9\% \pm 4.5\%$ (Table S9). Notably, cg12876356 exhibited the largest impact, with differences between replicates as high as 62 %. Despite the indication that correction of methylation data might not provide more accurate or reproducible results, for the sake of completeness, we decided to perform our downstream analysis on both the uncorrected and corrected datasets.

3.3. Smoking DNA methylation signatures in current, former and never smokers

Tobacco smoking exposure has a substantial impact on genome-wide DNA methylation levels of white blood cells in active smokers as it has been demonstrated via multiple large-scale EWAS [30]. Here, using our novel MPS assay, we analyzed the methylation status of 13 previously established smoking-associated and predictive CpGs in 232 whole blood samples with known smoking-related phenotypes, including 90 current smokers, 71 former smokers and 71 never smokers (Tables S10–11). Hierarchical clustering based on the uncorrected methylation values across CpGs / samples, revealed smoking-associated signatures for these three categories (Fig. 4A). In particular, the vast majority of current and never samples clustered separately; nevertheless, this was more challenging for former smokers.

We found evidence of methylation ‘transition’ from current to former to never smokers (or vice versa) for most CpGs (Fig. 4B), with the direction of methylation change being highly concordant with the ones obtained from microarray analysis presented by Maas et al. Similarly, when looking at the corrected methylation values across CpGs / samples, we observed similar signatures and directionality in methylation levels, with cg12876356 showing the largest PCR amplification bias (Fig. S6A–B). Furthermore, we uncovered statistically significant differences in methylation patterns for most analysed smoking-CpGs (except cg12803068, cg06126421 and cg09935388) driven by technology-related bias (Fig. S7A, Table S12), with PCR bias correction

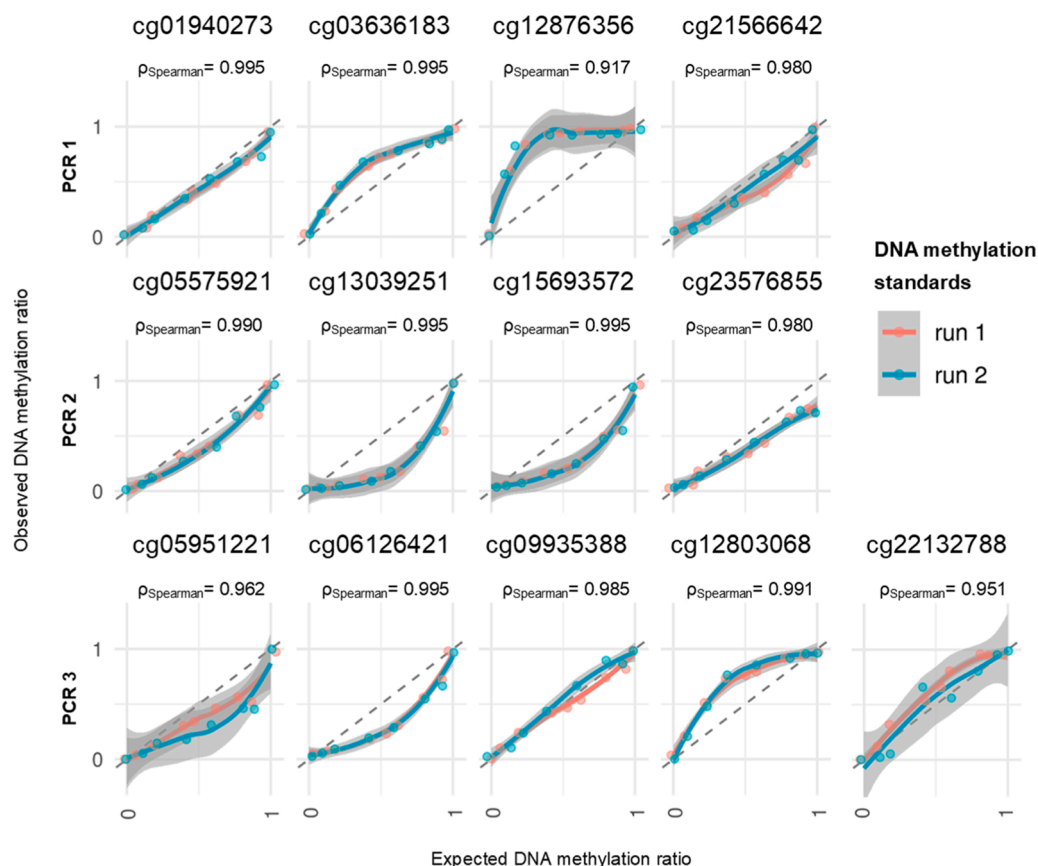


Fig. 3. Calibration curves obtained by analysing artificially methylated DNA standards (0, 10, 20, 40, 50, 60, 80, 90, 100 %) per CpG per MPS run. CpGs are segregated per multiplex PCR reaction. Fitted curves correspond to a locally estimated scatterplot smoothing (loess) fit. ρ_{spearman} : combined Spearman correlation coefficient across runs.

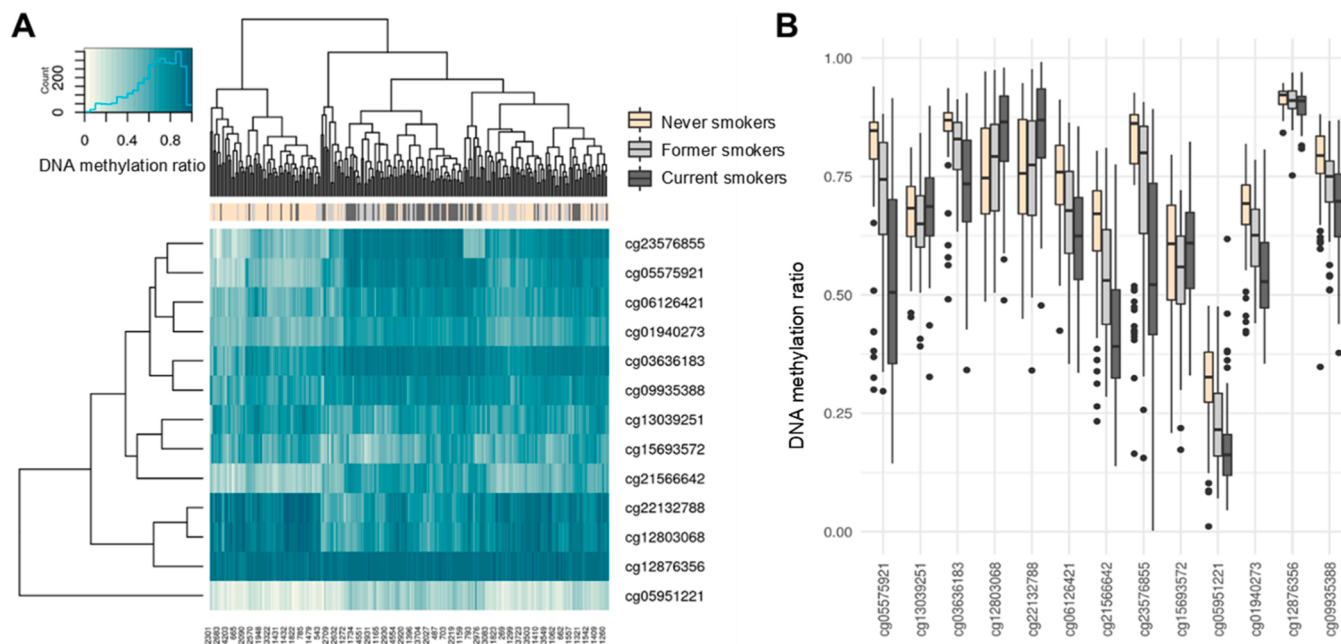


Fig. 4. DNA methylation relationship with smoking habits (current, former, never smokers). (A) Heatmap and dendrogram on the methylation ratio values across CpGs (rows, $n = 13$) and samples (columns, $n = 232$), where the upper row color indicates smoking habits; (B) Boxplot representing the methylation distribution per CpG between smoking categories. DNA methylation signals without PCR bias correction were employed for all visualizations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

failing to harmonize microarray and MPS data (Fig. S7B). Additionally, large inter-individual variation within each smoking category was observed for most smoking-CpGs (Fig. 4B), but particularly in former and current smokers, potentially reflecting the complex interplay in smoking-related behaviours (smoking intensity, duration, cigarette brand, etc). Hence, we also tested the correlation of methylation levels with smoking-related traits, such as the number of cigarettes smoked per day in current smokers and time since smoking cessation in former smokers. The vast majority (10 out of 13) of smoking-CpGs revealed a strong statistically significant correlation with daily cigarettes (p_{Pearson} ranging from -0.542 to 0.281 , Bonferroni-corrected p -value $< 0.001/13$) (Fig. 5A, Table S13); except for cg12876356, cg13039251 and cg15693572. Interestingly, regarding time since cessation, we only revealed a weak statistically significant association for cg15693572 (Fig. 5B).

3.4. Additional features of the 13-CpG marker pool

Smoking-related DNA methylation has been previously associated with epigenetic age acceleration [58], particularly in lungs [59], which additionally revealed sex-specific effects that may explain why women appear more susceptible to cigarette smoke and often develop more severe lung diseases [60]. Hence, we first investigated potential additional ageing and sex effects, using our diverse dataset (108 males / 124 females, aged from 22 to 82 years). As a result, eight smoking-CpGs revealed a statistically significant negative association with chronological age (p_{Pearson} ranging from -0.398 to -0.191 , p -value $< 0.05/13$), even after correcting for smoking habits (Fig. 6, Table S14). Additionally, when comparing sex-specific methylation levels of all 13 smoking-CpGs, only cg09935388 revealed a statistically significant association (p -value $= 1.46 \times 10^{-4}$); more specifically, mean methylation levels were lower in females (Fig. S8A).

Overall, we also validated co-methylation among smoking-CpGs (Fig. S8B). This is the result of the influence of common sources of variation including smoking and age, but also chromosomal proximity for those amplicons that happen to be nearby. Finally, we unraveled strong genetic effects on the methylation levels of cg23576855, as shown in both microarray and MPS data (Fig. S9A), driven by a common SNP (rs6869832) located on the CpG site (Fig. S9B) that disguises itself as the unmethylated allele following bisulfite conversion (Fig. S9C).

3.5. Prediction of smoking habits using previous and new statistical models

Our ultimate aim was to predict smoking habits from the generated MPS-based methylation data in blood. First, we tested the previously published, microarray-based statistical models from Maas et al.: a two-category multivariate logistic regression model (current versus non-current smoker) and a three-category multivariate multinomial logistic regression model (current versus former versus never smoker) on our MPS data [40]. In the original microarray data, we obtained an F_1 -score and a macro-weighted F-score of 0.956 and 0.709 (on the training set), for the two- and three-category model, respectively, where values of 1 correspond to perfect assignment. We place the F_1 -score baseline (i.e. random assignment with probability equal to the frequency of current smokers, r) at r ($511/3764 \approx 0.136$ and $90/232 \approx 0.388$ for the microarray and MPS data), respectively, and the macro-weighted F-score baseline (i.e. random assignment with probabilities equal to the frequency of current smokers, r_1 and former smokers, r_2) at $r_1^2 + r_2^2 + (1 - r_1 - r_2)^2$ (~ 0.400 and ~ 0.338 for the microarray and MPS data, respectively). When testing the models on our MPS data, these values corresponded to 0.675 and 0.649, respectively, with PCR bias correction worsening the performance of both models to a level just above the baseline (0.410 and 0.461, respectively). The prediction

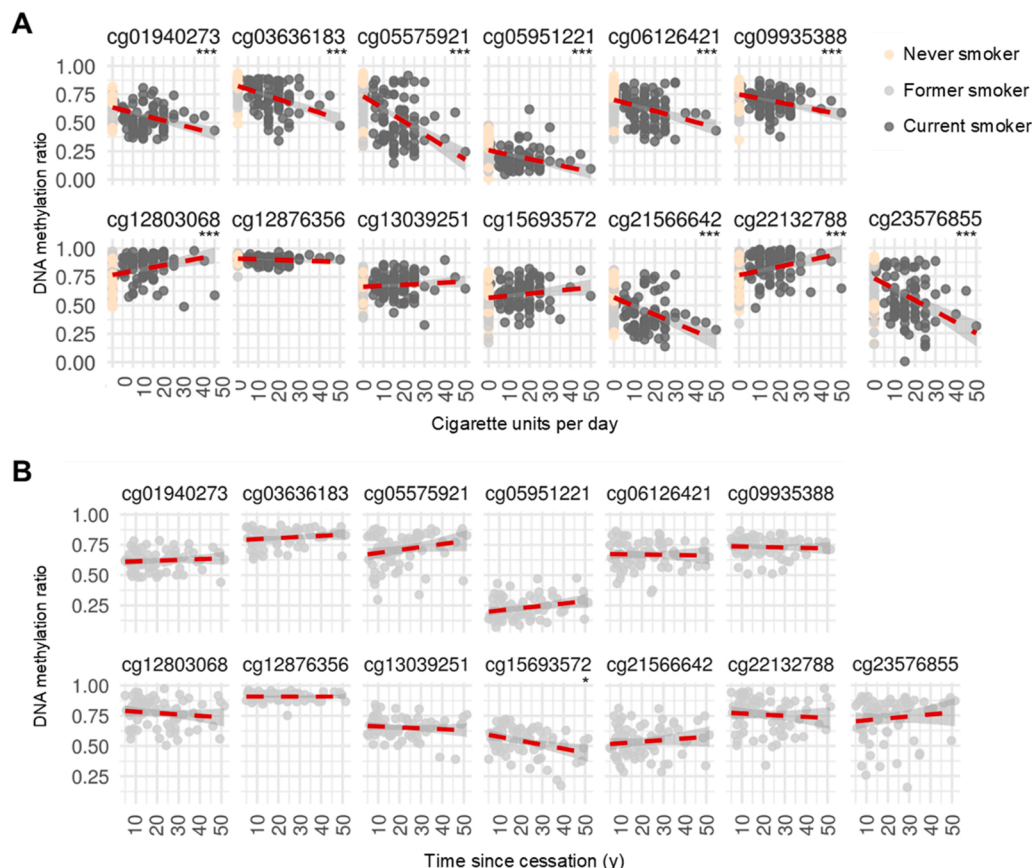


Fig. 5. DNA methylation relationship with smoking characteristics. Scatter plots on DNA methylation levels of each smoking-CpG against (A) the number of cigarette units smoked per day ($n = 229$, color-coded per smoking category); and (B) the time since cessation in former smokers ($n = 70$). Red fitted lines indicate ordinary least squares linear regression. DNA methylation signals without PCR bias correction were employed for all visualizations. *: p -val < 0.05 ; **: p -val < 0.01 ; ***: p -val < 0.001 . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

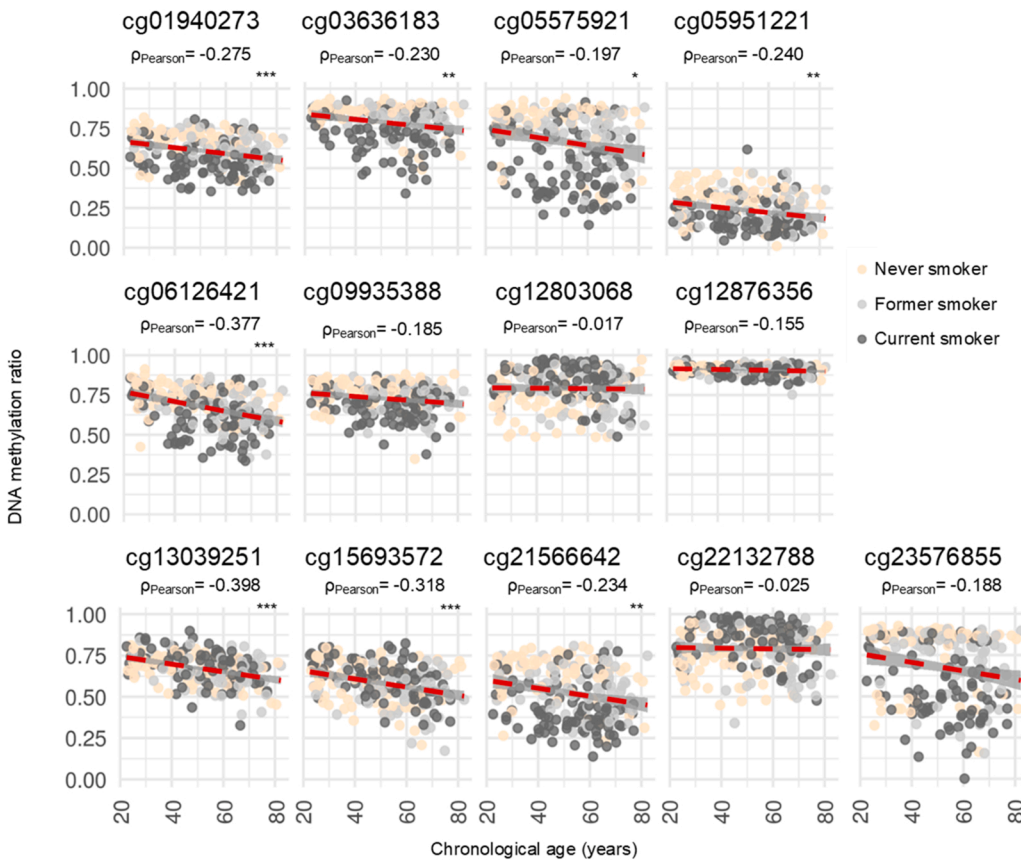


Fig. 6. Ageing influence on smoking-CpGs by plotting on DNA methylation levels of each smoking-CpG against chronological age ($n = 232$). Red fitted lines indicate ordinary least squares linear regression. DNA methylation signals without PCR bias correction were employed for all visualizations. ρ_{Pearson} : Pearson correlation coefficient; *: $p\text{-val} < 0.05$; **: $p\text{-val} < 0.01$; ***: $p\text{-val} < 0.001$; obtained via linear models and corrected for smoking status (three categories). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accuracy was substantially reduced in our MPS data, which can be explained by the technology-driven, statistically significant methylation differences we previously detected (Fig. S7, Tables S12–S13), but also the lack of validation of questionnaire information in our study.

Though the sample size for the MPS data was relatively smaller in comparison to microarray data, we aimed to account for technology-driven biases in model parameters. Additionally, we aimed to simultaneously inspect how these parameters change with respect to microarray data. One way of doing so is by building joint models for microarray and MPS data and introducing a covariate “technology” and interaction terms with all other predictors. This strategy is equivalent to building models on each technology separately and combining into a single model and allows side-by-side comparison of the obtained parameters. In total, we trained four new models: microarray + raw MPS data and microarray + corrected MPS data, for both two- and three-category smoking prediction and tested their performance with a repeated 5-

fold cross-validation (CV_{5-2}). As shown in Table 2, the obtained prediction accuracy for the microarray data subset was not affected by the inclusion of either raw or corrected MPS data in the two- and three-category models. Briefly, in the two-category model, 3174 out of 3253 (97.6 %) non-current smokers and 299 out of 511 (58.5 %) current smokers were correctly assigned (Fig. 7A), while for the three-category model, correct predictions corresponded to 78 % (969 out of 1243), 65.2 % (868 out of 1332) and 66.8 % (243 out of 364) of the never, former and current smokers, respectively (Fig. 7B).

For the MPS data subset, however, the obtained prediction accuracy was improved substantially relative to microarray-based only models. Briefly, in the two-category model, 123 out of 142 (86.6 %) of non-current smokers and 58 out of 90 (64.4 %) of current smokers were correctly assigned (Fig. 7A), while for the three-category model, correct predictions correspond to 73.2 %, 50.7 % and 71.1 % for the never, former and current smokers, respectively (Fig. 7B). The prediction accuracy of the most forensically relevant current smoking category between the two models was comparable. Interestingly, when looking at the probability space – the probabilities of each sample belonging to each category – we observed a lack of samples between never and current smokers, given the expected transition through former smokers (Fig. 7C). When comparing the outcomes of both models, there is high concordance in systematic sample misassignment (Fig. S8C). There was high agreement in prediction outcomes by the different-category models also for the technical replicates (Fig. 7D). Finally, similar performance was obtained for the corrected MPS subset (Table 2).

4. Discussion

Forensic DNA methylation profiling is a relatively new, but fast-developing field [6]. While so far it has mainly been used for identifying forensically relevant tissues and estimating chronological age, it holds great promises for the prediction of other (externally visible)

Table 2

Performance metrics (cross-validation per type of methylation data) of the newly trained 2- and 3-category, joint microarray:MPS smoking prediction models.

Joint model	Performance metric	Microarray data		MPS data	
		when trained with MPS data			
		Unprocessed	PCR bias Corrected	Unprocessed	PCR bias Corrected
2-cat	Average F_1 -score ($CV_{5,2}$)	0.956 ± 0.003	0.956 ± 0.005	0.809 ± 0.019	0.803 ± 0.052
3-cat	Average macro-weighted F-score ($CV_{5,2}$)	0.701 ± 0.021	0.702 ± 0.009	0.610 ± 0.068	0.609 ± 0.078

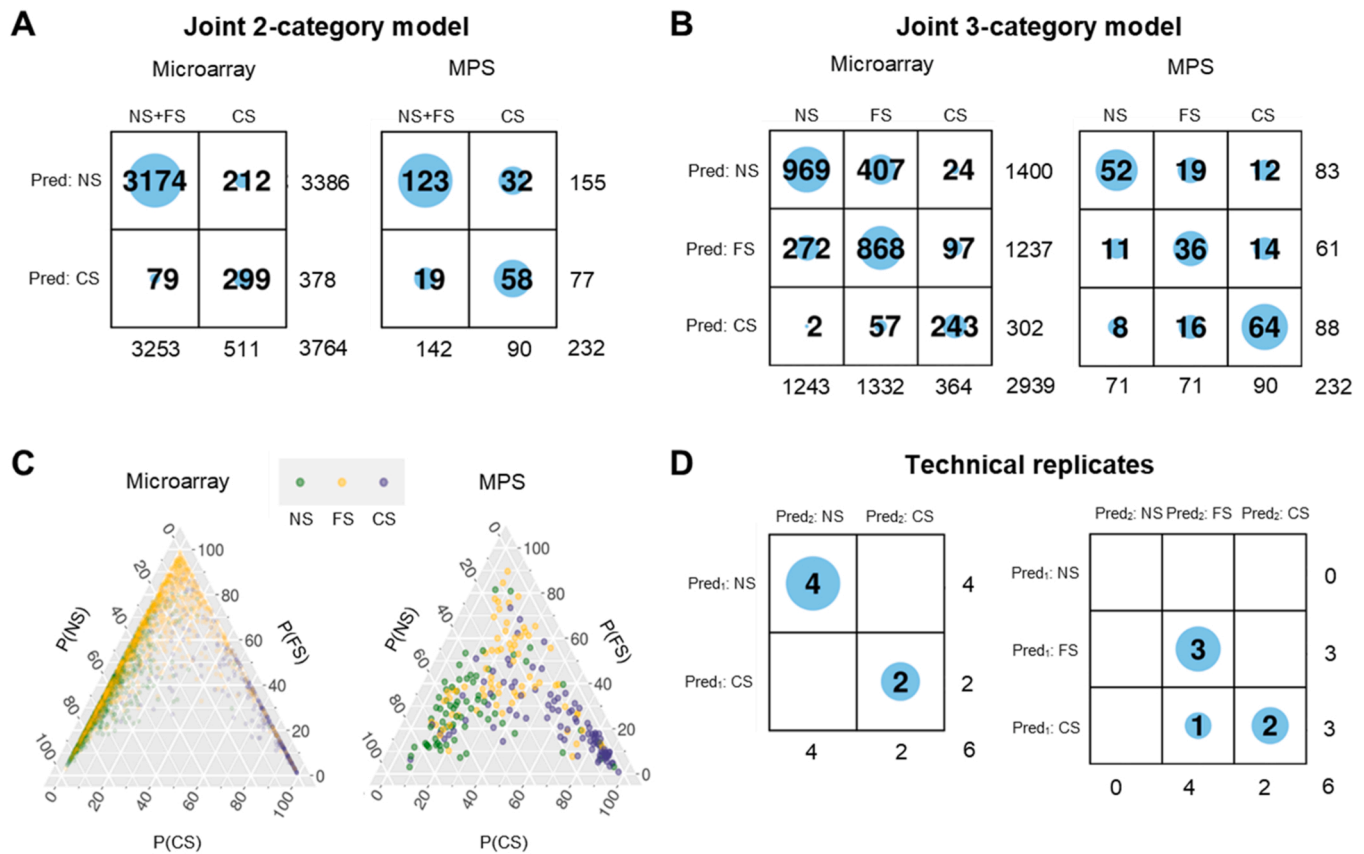


Fig. 7. Prediction of smoking habits from DNA methylation. (A) 2×2 and (B) 3×3 confusion matrices obtained for 2- or 3-category joint microarray:MPS models segregated by technology; (C) Ternary plots on the probabilities obtained by the 3-category joint model segregated by technology; (D) Agreement in the prediction output between technical duplicates by the 2- and 3-category joint models. MPS data is without PCR bias correction. NS: never smokers; FS: former smokers; CS: current smokers.

characteristics that can broaden investigative intelligence, such as lifestyle traits determined by environmental exposure [15]. Tobacco smoking is not only among the most frequent of such lifestyle habits (1 out of 5 adults smokes worldwide), but also among the most studied environmental modifiers in the human epigenome [61]. Genome-wide DNA methylation profiling studies have discovered thousands of smoking-associated CpGs in whole blood [30]. In this study, our main goal was to develop a novel MPS-based methylation assay to investigate blood methylation signatures of 13 previously reported smoking-associated CpGs and to assess both published and newly fitted models in their ability to predict smoking habits from the newly generated MPS data.

While there are previous small-scale attempts to develop targeted smoking prediction assays [62–65], to the best of our knowledge, our study offers the first-of-its-kind MPS assay to do so, which is a more appropriate technology for a future forensic application. Specifically, Shenker et al. used bisulfite pyrosequencing to analyze four genomic loci, which they used for calculating a single methylation index for smoking habit prediction [62]. While pyrosequencing has been used for other forensic epigenetic assays before [6], it is mainly employed in single-plex form which is not advantageous for future applications. Additionally, it includes a high number of PCR cycles (usually 45) that could further enhance amplification bias. Similarly, Philibert et al. employed digital PCR to analyze a single smoking CpG (cg05575921) and predict smoking habits for diagnostic purposes [63]. Despite the promising higher resolution of digital PCR in terms of methylation detection, this approach also suffers from low multiplex capability. Additionally, we do not favour single-CpG models. Kondratyev et al. aimed to account for this issue by employing single-molecule, long-read

sequencing, which gave the added benefit of studying allele-specific methylation effects [65]. Nevertheless, PacBio sequencing is expensive and its true forensically relevant benefits against Illumina sequencing when analyzing only five targets that the study employed, is still to be determined. Finally, more recently, Wen et al. proposed a new assay based on methylation-sensitive single-nucleotide primer extension to study nine genomic loci for smoking habit prediction [64]. This approach has also been used for other forensic applications [6], and it offers methodological benefits compared to MPS, particularly when analyzing a small number of samples.

Over the last years, targeted MPS is gaining substantial attention in forensic genetics [66], as it offers compelling benefits over standard detection techniques, particularly on combined sensitivity, multiplex capability and resolution. As a result, more recently, researchers developed MPS assays for targeted methylation detection as part of various applications [10,11,56]. Nevertheless, there is a range of epigenetics-specific issues that we encountered, particularly during multiplex PCR design, amplification and sequencing. First of all, while bisulfite conversion is currently the golden standard in epigenetic assays, it is clear that it is not suited for forensic typing. It not only requires a large amount of genomic DNA input (optimal: 200 – 500 ng), often unavailable from forensic-type material, but can also further degrade the often already degraded DNA. In our experiments, due to their pilot nature, we decided to ignore this problem and opted for optimal input DNA amounts per sample (200 ng), as at this stage our priority was to produce high-quality data. Future work should further explore forensically relevant sensitivity of bisulfite conversion and additionally consider novel alternative, non-chemical conversion techniques, such as enzymatic-based, that promise to offer more sensitive and gentle

detection [67].

Due to its short, single-stranded and reduced-complex (T-rich) nature, converted DNA is particularly difficult to amplify, which we naturally also observed in this study. Not only did we encounter problems with low primer specificity and PCR assay efficiency at the single-plex PCR level, but it turned out impossible to co-amplify 13 PCR targets in a single multiplex reaction. Future strategies, for example via novel DNA polymerases or probe capture, might solve these issues and allow for increasing the number of CpGs we can simultaneously detect in targeted epigenetic assays. Nevertheless, we managed to develop three small-scale multiplex PCR reactions, which, compared to single-plex reactions, allowed us to increase sensitivity, cost-effectiveness and scalability. Despite opting for 10 ng of converted DNA PCR input towards reducing the number of amplification cycles, we still detected substantial amplification bias in artificially methylated DNA controls for most assays. Such amplification bias is known in the field [57] and reported in recent MPS-based forensic epigenetic assays for age prediction [68]. To account for these biases, we followed a novel correction approach based on bi-exponential models, which, to the best of our knowledge, have not been employed for such purpose before. In this context, bi-exponential models tend to be more appropriate for obtaining monotonic fits than third-degree polynomials (typical choice), since the latter exploit their larger degrees of freedom to generate unphysical crests and troughs that pass through the minutiae of the data (overfitting). Despite this, the correction strategy did not result in improved accuracy of methylation detection, nor follow-up prediction accuracy of smoking habits. The success of a PCR amplification correction strategy heavily depends on the level of the observed methylation per marker and also depends on the contribution of each CpG to the prediction model. One could argue that primer redesign could solve the issue of amplification bias, but this can be particularly challenging for CpG-dense fragments (~one CpGs every ten bases). In these cases, we are forced to include degenerate bases in primer sequences to avoid preferential binding, but it seems that this likely affects methylation detection accuracy, such as in assays cg12876356 and cg05951221 in our study. Alternatively, a reduced PCR cycle (i.e., 30–32 cycles) may be employed to minimize effects across assays, as long as the required input for library preparation is still met. Future work should explore additional correction strategies, for example by using UMIs to track amplification events, as recently proposed for PCR bias in other (RNA-based) assays [69].

For sequencing, and to avoid further amplification bias, we employed an adjusted version of an amplification-free library preparation protocol and used the maximum possible PCR product input (500 ng). This resulted in sufficient, relatively balanced sequencing coverage across amplicons, with an average of ~15,000 reads per CpG that resulted in reproducible methylation detection. Our detected mean standard deviation between technical replicates across assays matched the ones reported also by other recent studies [68]. Generally, obtained sequencing depth corresponded somewhat with expectations from PCR amplification efficiencies; for example, we did obtain the faintest band for cg21566642, with the longest amplicon (264 bp). However, this was not the case for cg09935388, which was the shortest amplicon in our assay (125 bp), indicating potential post-PCR analysis bias like preferential binding during bead clean-up. Future research is required to determine the minimum read coverage per CpG required for the purpose of smoking prediction, that would allow for co-analyzing a larger number of samples to reduce costs. Full developmental validation of the MPS assay is also needed, especially in terms of sensitivity and robustness, to be able to make conclusions on forensic applicability in the future.

Using our novel MPS method, we examined the methylation of the selected 13 smoking-CpGs in the blood of 232 Europeans, classified as current ($n = 90$), former ($n = 71$) and never ($n = 71$) smokers. While there was a large inter-individual variation in methylation levels within each category, we were able to detect distinct smoking-associated

signatures in former and current smokers, where most smoking-CpGs were becoming less methylated upon smoking exposure. Overall, there was consistency between the detected microarray and MPS methylation, but also clear technology-driven differences for certain CpGs, as expected. We did not detect Bonferroni-significant association between methylation levels and number of daily cigarettes for all 13 smoking CpGs, but this is likely driven by low power as only a subset of samples had available phenotypes and the highly conservative nature of Bonferroni multiple testing correction approach. Specifically, for the strongest smoking-associated CpGs in our study (cg05575921), we detected almost identical methylation levels between non- and current smokers as previously reported in the literature [65,70,71], even if these were measured by different technologies. We also confirmed its strong association with smoking intensity, closely linked to lung cancer risk and mortality [72]. Nevertheless, understanding baseline methylation of smoking-CpGs across populations is important to be able to fully characterize the smoking-induced signatures in future studies. For example, methylation at cg05575921 was reported lower in non-smoking adults residing in areas with high air pollutants (fine particulate matter, $PM_{2.5}$) [73,74]. Finally, in our study we only investigated the smoking association of the selected 13 smoking-CpGs, purposefully ‘ignoring’ other CpGs included in the same amplicons. It may be beneficial to investigate adjacent CpGs in follow-up studies, especially to replace CpGs with observed weak smoking correlations, taking the full advantage of MPS analysis.

Moreover, while we classified two or three distinct smoking categories, we recognize that smoking is a rather quantitative trait. Indeed, methylation levels of the vast majority of the smoking-CpGs included in this study showed significant association with the number of cigarettes smoked per day, similar to other studies [34,35,75], but not with time since cessation, likely due to our reduced power. For example, McCartney et al. found that the exposure time point from which at least 50 % of current smokers were assigned to the smoker-enriched clustered varied between 5 and 9 years in heavier smokers, and between 15 and 19 years in lighter smokers [35]. Additionally, low-dose former smokers were more likely to be assigned to the never smoker-enriched cluster in the first year following cessation, in contrast to two years for the high-dose former smokers [35]. In the future, in-depth longitudinal analysis will allow us to recognize how universal and reversible dose-dependent smoking-associated effects are. For example, Philibert et al. tested whether methylation at cg05575921 can be used to verify cessation by determining monthly levels in smokers undergoing biochemically monitored contingency management-based smoking cessation therapy [76]. They found that cg05575921 methylation reversion was dependent on their initial smoking intensity, with methylation levels in the heaviest smokers reverting to an average of 0.12 % per day over a 3-month period [76]. Dedicated large-scale studies to assess reversibility will shed light on the dynamic nature of smoking-associated DNA methylation [77]. In the forensic context, we understand that a two-categorical prediction of current versus non-current smokers may be more suitable as the focus is on the current habits; nevertheless, it is important to first understand all smoking-related behaviours including in former smokers. In the recent study by Wen et al., the authors completely excluded former smokers from their predictive analysis, which heavily impacts their conclusions and does not allow to compare their outcomes with our study, despite the six overlapping CpGs [64].

In this study, we focused on data from Europeans, because the smoking-associated CpGs we used here were previously established in Europeans. However, smoking association of DNA methylation in blood, including its dose dependence, may be biogeographical ancestry-specific, as shown by an increasing number of studies including in African Americans and South Asians [75,78,79]. Future efforts should focus to discover biomarkers that better capture smoking effects across diverse human populations. Also, in our study, we focused on blood, as the employed predictive markers were previously identified in blood. Although blood represents the most studied tissue source so far for

smoking-associated DNA methylation, it is a complex tissue containing different cell types that may have different DNA methylation patterns. In the future, it would be interesting to study how differences in cell type composition affects the prediction of smoking habits by extended cell type-specific methylation profiling. In fact, it was recently proposed that most of the highly reproducible smoking-related hypomethylation signatures are more prominent in the myeloid lineage, compared to other immune-cell subtypes [80]. Additionally, it is worth investigating other forensically relevant tissues, such as saliva [70,71], sperm [81,82] or brain [83,84], with reported similar smoking-associated effects. Particularly, for cg05575921 both baseline and smoking-related methylation levels were lower in saliva compared to blood [71]. Nevertheless, no such investigation has been performed for other more frequently forensically encountered tissues, such as skin.

Importantly, in our study, the vast majority of smoking-CpGs were found strongly associated with age. While the literature suggests that active smoking exposure accelerates DNA methylation age in blood [58, 85,86] in a quantitative way associated to pack-years [87], here we found that age is associated to methylation reduction in smoking-associated CpGs even after correcting for smoking habits. In the forensic scenario, this is relevant as smoking prediction might additionally reveal information on the donor's age, and *vice versa*. Nevertheless, the age predictive value of the employed smoking-CpGs is still to be determined. We also detected a sex-associated methylation of a single CpG, adding upon other CpGs in the literature [65]. This highlights the complexity, inter-relationship and effects of genetic and environmental factors on DNA methylation, which should be fully understood before predictions based on such markers are used in practice. Furthermore, exploring the association of smoking with other traits such as body mass index would be beneficial [88] to improve our understanding on complex traits and ensure successful (legal) practical application, such as in forensics.

Towards our ultimate goal, we employed the generated MPS data to predict smoking habits from the observed DNA methylation patterns. For this, we first tested previously published models by Maas et al. based on microarray data and secondly, we established new models by combining microarray and MPS data. Despite that three out of the 13 smoking CpGs were not significantly associated with daily cigarettes, we decided to include all markers for this analysis, to provide with fair performance comparisons side-by-side with the previously published models. As expected, training joint models that include MPS data improved accuracy for both two- and three-category prediction for MPS data by allowing inter-technological corrections. However, the false negative rate is still relatively high, which has been reported before in single-smoking CpG models [62]. Misclassifications may result from the effects of various factors. Firstly, the 13 CpGs included in our study capture only a small portion of the epigenome-wide smoking signature. Including more markers, will allow us to gain more resolution and improved predictions; but of course, the number of markers to be analyzed fully depends on the employed technology. Secondly, given the quantitative nature of smoking and the diversity of traits within each (non-)smoker, it might be more suitable to envision this problem as a regression problem rather than a classification problem; for example, by predicting a smoking exposure-related, quantitative variable, like number of cigarettes per day. Thirdly, without having extra biological data confirming smoking habits, such as cotinine levels, we are fully dependent on the information of the self-reported questionnaires, which are known to contain discrepancies [22]. At the same time, we cannot exclude the effects of passive smoking, which could potentially explain why some non-smokers are predicted as smokers. Future work may also investigate the effect of other smoking traits or behaviors on DNA methylation, such as vaping and non-combustible tobacco use [89], to name but a few. Moreover, limited access to suitable samples and resources prevented us from increasing the sample size in this study. Hence, the analysis of more samples based on the above-mentioned conditions will allow us to retrain MPS-only models, including via

other modelling approaches, with the potential of improved performance. Additionally, smoking habit predictions can be corrected with other traits (such as sex and age) for improved accuracy and combined for a more complete picture of a personalized epigenomic fingerprint. As a final point, efforts to discover and combine other types of molecular biomarkers of smoking, such as single nucleotide polymorphisms [90–92], RNA markers [84,93] and microbial DNA [94], are worth exploring.

5. Conclusion

In this study, we developed a novel MPS assay to investigate smoking methylation signatures in blood, particularly of 13 smoking-associated CpGs included in Maas et al. Assessment of the method with technical duplicates and artificially methylated controls revealed that it is accurate and reproducible, but it also suffers from known amplification bias. We applied this method to a relatively large set of blood sample from a population cohort and examined methylation differences between current, former and never smokers, and associated methylation patterns with smoking-related behaviours (daily cigarettes and time since cessation) as well as with other traits such as age and sex. We found that the prediction accuracy for both two- and three-category models was sustained by retraining previous microarray data with newly generated MPS data, to account for technology-driven variation. In summary, our method can be used to further study smoking-associated methylation patterns in blood and other tissues and brings us a step closer to the future forensic application.

6. CRediT authorship contribution statement

Athina Vidaki: Conceptualization, Methodology, Formal analysis, Visualization, Supervision, Writing – Original draft; **Benjamin Planterose Jiménez:** Methodology, Software, Formal analysis, Data curation, Visualization, Writing – Original draft; **Brando Poggiali:** Validation, Investigation, Writing – Original draft; **Vivian Kalamara:** Methodology, Investigation, Writing - Review & Editing; **Kristiaan van der Gaag:** Investigation, Writing - Review & Editing; **Silvana Maas:** Methodology, Writing - Review & Editing; **Mohsen Ghanbari:** Resources; **Titia Sijen:** Resources, Writing - Review & Editing; **Manfred Kayser:** Supervision, Resources, Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Raw and corrected DNA methylation data as well as information on 2- and 3-category models are available for replication in the [Supporting Information \(Tables S4–5, S7–11\)](#). Data analysis was performed by employing custom R-scripts and models, which have been released to the public domain under an MIT licence at GitHub (<https://github.com/BenjaminPlanterose/GAMBA>) and at the Zenodo digital object identifier-assigning repository (<https://zenodo.org/record/7404631#.Y48fhzPMI3w>).

Acknowledgements

We would like to thank the donors from the Erasmus Rucphen Family (ERF) study that have donated whole blood samples included in this study. We are also grateful to Ivana Prokic (Dept. Epidemiology, Erasmus MC) and Arwin Ralf (Dept. Genetic Identification, Erasmus MC) for their technical assistance with metadata and sample curation, respectively. Methylation microarray data for the 13 smoking-CpGs were

kindly provided by six Dutch cohorts embedded within the Biobank-based Integrative Omics Study (BIOS) Consortium: LifeLines, the Leiden Longevity Study, the Netherlands Twin Registry (NTR), the Rotterdam Study, the Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) study, and the Prospective ALS study Netherlands (PAN). We would like to thank the participants of all aforementioned biobanks and their investigators. This research was financially supported by Erasmus MC and the Netherlands Forensic Institute. AV was additionally supported by an Erasmus MC fellowship 2020.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102878.

References

- [1] J.M. Butler, Short tandem repeat typing technologies used in human identity testing, *Biotechniques* 43 (4) (2007) ii–v, <https://doi.org/10.2144/000112582>.
- [2] M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, *Forensic Sci. Int. Genet.* 18 (2015) 33–48, <https://doi.org/10.1016/j.fsigen.2015.02.003>.
- [3] P.M. Schneider, B. Prainsack, M. Kayser, The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry, *Dtsch Arztebl. Int.* 51–52 (2019) 873–880, <https://doi.org/10.3238/arztebl.2019.0873>.
- [4] C. Xavier, M. de la Puente, A. Mosquera-Miguel, A. Freire-Aradas, V. Kalamara, A. Ralf, A. Revoir, T.E. Gross, P.M. Schneider, C. Ames, C. Hohoff, C. Phillips, M. Kayser, W. Parson, Development and inter-laboratory evaluation of the VISAGE enhanced tool for appearance and ancestry inference from DNA, *Forensic Sci. Int. Genet.* 61 (2022), 102779, <https://doi.org/10.1016/j.fsigen.2022.102779>.
- [5] A. Vidaki, B. Daniel, D.S. Court, Forensic DNA methylation profiling—potential opportunities and challenges, *Forensic Sci. Int. Genet.* 7 (5) (2013) 499–507, <https://doi.org/10.1016/j.fsigen.2013.05.004>.
- [6] A. Vidaki, M. Kayser, Recent progress, methods and perspectives in forensic epigenetics, *Forensic Sci. Int. Genet.* 37 (2018) 180–195, <https://doi.org/10.1016/j.fsigen.2018.08.008>.
- [7] G.A. Dhar, S. Saha, P. Mitra, R. Nag Chaudhuri, DNA methylation and regulation of gene expression: guardian of our health, *Nucleus* 64 (3) (2021) 259–270, <https://doi.org/10.1007/s13237-021-00367-y>.
- [8] 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, <https://doi.org/10.1016/j.fsigen.2016.06.007>.
- [9] A. Vidaki, F. Giangasparo, D. Syndercombe Court, Discovery of potential DNA methylation markers for forensic tissue identification using bisulphite pyrosequencing, *Electrophoresis* 37 (21) (2016) 2767–2779, <https://doi.org/10.1002/elps.201600261>.
- [10] 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, <https://doi.org/10.1016/j.fsigen.2018.09.003>.
- [11] J. Naue, T. Sängner, H.C.J. Hoefsloot, S. Lutz-Bonengel, A.D. Kloosterman, P. J. Verschure, Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing, *Forensic Sci. Int. Genet.* 36 (2018) 152–159, <https://doi.org/10.1016/j.fsigen.2018.07.007>.
- [12] 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, <https://doi.org/10.1016/j.fsigen.2017.04.006>.
- [13] A. Vidaki, C. Díez López, E. Carnero-Montoro, A. Ralf, K. Ward, T. Spector, J. T. Bell, M. Kayser, Epigenetic discrimination of identical twins from blood under the forensic scenario, *Forensic Sci. Int. Genet.* 31 (2017) 67–80, <https://doi.org/10.1016/j.fsigen.2017.07.014>.
- [14] A. Vidaki, V. Kalamara, E. Carnero-Montoro, T.D. Spector, J.T. Bell, M. Kayser, Investigating the epigenetic discrimination of identical twins using buccal swabs, saliva, and cigarette butts in the forensic setting, *Genes* 9 (5) (2018), <https://doi.org/10.3390/genes9050252>.
- [15] A. Vidaki, M. Kayser, From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence, *Genome Biol.* 18 (1) (2017) 238, <https://doi.org/10.1186/s13059-017-1373-1>.
- [16] K. Kim, Y. Zheng, B.T. Joyce, H. Jiang, P. Greenland, D.R. Jacobs Jr., K. Zhang, L. Liu, N.B. Allen, J.T. Wilkins, S.N. Forrester, D.M. Lloyd-Jones, L. Hou, Relative contributions of six lifestyle- and health-related exposures to epigenetic aging: the Coronary Artery Risk Development in Young Adults (CARDIA) Study, *Clin. Epigenet.* 14 (1) (2022) 85, <https://doi.org/10.1186/s13148-022-01304-9>.
- [17] WHO, World Health Organization Fact Sheet: Tobacco, 2022. <https://www.who.int/news-room/fact-sheets/detail/tobacco>. (Accessed 15 November 2022).
- [18] X. Dai, G.F. Gil, M.B. Reitsma, N.S. Ahmad, J.A. Anderson, C. Bisignano, S. Carr, R. Feldman, S.I. Hay, J. He, V. Iannucci, H.R. Lawlor, M.J. Malloy, L.B. Marczak, S. A. McLaughlin, L. Morikawa, E.C. Mullany, S.I. Nicholson, E.M. O'Connell, C. Okereke, R.J.D. Sorensen, J. Whisnant, A.Y. Aravkin, P. Zheng, C.J.L. Murray, E. Gakidou, Health effects associated with smoking: a burden of proof study, *Nat. Med.* 28 (10) (2022) 2045–2055, <https://doi.org/10.1038/s41591-022-01978-x>.
- [19] X. Gao, N. Huang, M. Jiang, B. Holleczech, B. Schottker, T. Huang, H. Brenner, Mortality and morbidity risk prediction for older former smokers based on a score of smoking history: evidence from UK Biobank and ESTHER cohorts, *Age Ageing* 51 (7) (2022), <https://doi.org/10.1093/ageing/afac154>.
- [20] Eurostat, Tobacco Consumption Statistics, 2019. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Tobacco_consumption_statistics. (Accessed 15 November 2022).
- [21] Z. Herceg, S. Ambatipudi, Smoking-associated DNA methylation changes: no smoke without fire, *Epigenomics* 11 (10) (2019) 1117–1119, <https://doi.org/10.2217/epi-2019-0136>.
- [22] S. Connor Gorber, S. Schofield-Hurwitz, J. Hardt, G. Levasseur, M. Tremblay, The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status, *Nicotine Tob. Res.* 11 (1) (2009) 12–24, <https://doi.org/10.1093/ntr/ntn010>.
- [23] D. Shipton, D.M. Tappin, T. Vadiveloo, J.A. Crossley, D.A. Aitken, J. Chalmers, Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study, *BMJ* 339 (2009) b4347, <https://doi.org/10.1136/bmj.b4347>.
- [24] N.L. Benowitz, Cotinine as a biomarker of environmental tobacco smoke exposure, *Epidemiol. Rev.* 18 (2) (1996) 188–204, <https://doi.org/10.1093/oxfordjournals.epirev.a017925>.
- [25] M. Fricker, B.J. Goggins, S. Mateer, B. Jones, R.Y. Kim, S.L. Gellatly, A.G. Jarnicki, N. Powell, B.G. Oliver, G. Radford-Smith, N.J. Talley, M.M. Walker, S. Keely, P. M. Hansbro, Chronic cigarette smoke exposure induces systemic hypoxia that drives intestinal dysfunction, *JCI Insight* 3 (3) (2018), <https://doi.org/10.1172/jci.insight.94040>.
- [26] N.H. Yamaguchi, Smoking, immunity, and DNA damage, *Transl. Lung Cancer Res.* 8 (Suppl 1) (2019) S3–S6, <https://doi.org/10.21037/tlcr.2019.03.02>.
- [27] J. Huang, M. Okuka, W. Lu, J.C. Tsibris, M.P. McLean, D.L. Keefe, L. Liu, Telomere shortening and DNA damage of embryonic stem cells induced by cigarette smoke, *Reprod. Toxicol.* 35 (2013) 89–95, <https://doi.org/10.1016/j.reprotox.2012.07.003>.
- [28] R. Satta, E. Maloku, A. Zhubi, F. Pibiri, M. Hajos, E. Costa, A. Guidotti, Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons, *Proc. Natl. Acad. Sci. USA* 105 (42) (2008) 16356–16361, <https://doi.org/10.1073/pnas.0808699105>.
- [29] 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, <https://doi.org/10.1016/j.ajhg.2011.03.003>.
- [30] 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. Epigenet.* 7 (2015) 113, <https://doi.org/10.1186/s13148-015-0148-3>.
- [31] S. Li, E.M. Wong, M. Bui, T.L. Nguyen, J.E. Joo, J. Stone, G.S. Dite, G.G. Giles, R. Saffery, M.C. Southey, J.L. Hopper, Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study, *Clin. Epigenet.* 10 (2018) 18, <https://doi.org/10.1186/s13148-018-0452-9>.
- [32] P.P. Mishra, I. Hanninen, E. Raitoharju, S. Marttila, B.H. Mishra, N. Mononen, M. Kahonen, M. Hurme, O. Raitakari, P. Toronen, L. Holm, T. Lehtimäki, Epigenome-450K-wide methylation signatures of active cigarette smoking: The Young Finns Study, *Biosci. Rep.* 40 (7) (2020), <https://doi.org/10.1042/BSR20200596>.
- [33] C. Christiansen, J.E. Castillo-Fernandez, A. Domingo-Relloso, W. Zhao, J.S. El-Sayed Moustafa, P.C. Tsai, J. Maddock, K. Haack, S.A. Cole, S.L.R. Kardina, M. Molokhia, M. Suderman, C. Power, C. Relton, A. Wong, D. Kuh, A. Goodman, K. S. Small, J.A. Smith, M. Tellez-Plaza, A. Navas-Acien, G.B. Ploubidis, R. Hardy, J. T. Bell, Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects, *Clin. Epigenet.* 13 (1) (2021) 36, <https://doi.org/10.1186/s13148-021-01018-4>.
- [34] R. Wilson, S. Wahl, L. Pfeiffer, C.K. Ward-Caviness, S. Kunze, A. Kretschmer, E. Reischl, A. Peters, C. Gieger, M. Waldenberger, The dynamics of smoking-related disturbed methylation: a two time-point study of methylation change in smokers, non-smokers and former smokers, *BMC Genom.* 18 (1) (2017) 805, <https://doi.org/10.1186/s12864-017-4198-0>.
- [35] D.L. McCartney, A.J. Stevenson, R.F. Hillary, R.M. Walker, M.L. Bermingham, S. W. Morris, T.K. Clarke, A. Campbell, A.D. Murray, H.C. Whalley, D.J. Porteous, P. M. Visscher, A.M. McIntosh, K.L. Evans, I.J. Deary, R.E. Marioni, Epigenetic signatures of starting and stopping smoking, *EBioMedicine* 37 (2018) 214–220, <https://doi.org/10.1016/j.ebiom.2018.10.051>.
- [36] K.A. McGinnis, A.C. Justice, J.P. Tate, H.R. Kranzler, H.A. Tindle, W.C. Becker, J. Concato, J. Gelernter, B. Li, X. Zhang, H. Zhao, K. Crothers, K. Xu, V.P. Group, Using DNA methylation to validate an electronic medical record phenotype for smoking, *Addict. Biol.* 24 (5) (2019) 1056–1065, <https://doi.org/10.1111/adb.12670>.
- [37] K. Sugden, E.J. Hannon, L. Arseneault, D.W. Belsky, J.M. Broadbent, D.L. Corcoran, R.J. Hancox, R.M. Houts, T.E. Moffitt, R. Poulton, J.A. Prinz, W.M. Thomson, B. S. Williams, C.C.Y. Wong, J. Mill, A. Caspi, Establishing a generalized polyepigenetic biomarker for tobacco smoking, *Transl. Psychiatry* 9 (1) (2019) 92, <https://doi.org/10.1038/s41398-019-0430-9>.
- [38] D.L. McCartney, R.F. Hillary, A.J. Stevenson, S.J. Ritchie, R.M. Walker, Q. Zhang, S.W. Morris, M.L. Bermingham, A. Campbell, A.D. Murray, H.C. Whalley, C. R. Gale, D.J. Porteous, C.S. Haley, A.F. McRae, N.R. Wray, P.M. Visscher, A. M. McIntosh, K.L. Evans, I.J. Deary, R.E. Marioni, Epigenetic prediction of complex

- traits and death, *Genome Biol.* 19 (1) (2018) 136, <https://doi.org/10.1186/s13059-018-1514-1>.
- [39] S. Bollepalli, T. Korhonen, J. Kaprio, S. Anders, M. Ollikainen, EpiSmoker: a robust classifier to determine smoking status from DNA methylation data, *Epigenomics* 11 (13) (2019) 1469–1486, <https://doi.org/10.2217/epi-2019-0206>.
- [40] S.C.E. Maas, A. Vidaki, R. Wilson, A. Teumer, F. Liu, J.B.J. van Meurs, A. G. Uitterlinden, D.I. Boomsma, E.J.C. de Geus, G. Willemsen, J. van Dongen, C.J. H. van der Kallen, P.E. Slagboom, M. Beekman, D. van Heemst, L.H. van den Berg, B. Consortium, L. Duijts, V.W.V. Jaddoe, K.H. Ladwig, S. Kunze, A. Peters, M. A. Ikram, H.J. Grabe, J.F. Felix, M. Waldenberger, O.H. Franco, M. Ghanbari, M. Kayser, Validated inference of smoking habits from blood with a finite DNA methylation marker set, *Eur. J. Epidemiol.* 34 (11) (2019) 1055–1074, <https://doi.org/10.1007/s10654-019-00555-w>.
- [41] A. Vidaki, Method for Determining Global Bisulfite Conversion Efficiency. <https://patentscope2.wipo.int/search/pt/detail.jsf?docId=WO2021048410>, 2019. (Accessed 25 April 2023).
- [42] B. Planterose Jiménez, M. Kayser, A. Vidaki, Revisiting genetic artifacts on DNA methylation microarrays exposes novel biological implications, *Genome Biol.* 22 (1) (2021) 274, <https://doi.org/10.1186/s13059-021-02484-y>.
- [43] UCSC, Lift Genome Annotations. <https://genome.ucsc.edu/cgi-bin/hgLiftOver>, 2022. (Accessed 16 November 2022).
- [44] Ensembl, GRCh37/hg19 genome browser. <http://grch37.ensembl.org/index.html>, 2022. (Accessed 15 December 2022).
- [45] L.C. Li, R. Dahiya, MethPrimer: designing primers for methylation PCRs, *Bioinformatics* 18 (11) (2002) 1427–1431, <https://doi.org/10.1093/bioinformatics/18.11.1427>.
- [46] T. Arányi, A. Váradi, I. Simon, G.E. Tusnády, The BiSearch web server, *BMC Bioinform.* 7 (1) (2006) 431, <https://doi.org/10.1186/1471-2105-7-431>.
- [47] P.M. Vallone, J.M. Butler, AutoDimer: a screening tool for primer-dimer and hairpin structures, *Biotechniques* 37 (2) (2004) 226–231, <https://doi.org/10.2144/04372ST03>.
- [48] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* 30 (15) (2014) 2114–2120, <https://doi.org/10.1093/bioinformatics/btu170>.
- [49] M. Martin, Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads, 2011. 17 (1) (2011) 3. <https://doi.org/10.14806/ej.17.1.200>.
- [50] S. Andrews, FastQC: a Quality Control tool for High Throughput Sequencing Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, 2010 (Accessed 11 November 2022).
- [51] P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, *Bioinformatics* 32 (19) (2016) 3047–3048, <https://doi.org/10.1093/bioinformatics/btw354>.
- [52] F. Krueger, S.R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics* 27 (11) (2011) 1571–1572, <https://doi.org/10.1093/bioinformatics/btr167>.
- [53] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The sequence alignment/map format and SAMtools, *Bioinformatics* 25 (16) (2009) 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352>.
- [54] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, *Nat. Biotechnol.* 29 (1) (2011) 24–26, <https://doi.org/10.1038/nbt.1754>.
- [55] R-Core-Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>, 2022. (Accessed 11 November 2022).
- [56] D.R. Masser, A.S. Berg, W.M. Freeman, Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing, *Epigenet. Chromatin* 6 (1) (2013) 33, <https://doi.org/10.1186/1756-8935-6-33>.
- [57] E.A. Moskalov, M.G. Zavgordnij, S.P. Majorova, I.A. Vorobjev, P. Jandaghi, I. V. Bure, J.D. Hoheisel, Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression, *Nucleic Acids Res.* 39 (11) (2011), e77, <https://doi.org/10.1093/nar/gkr213>.
- [58] 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, <https://doi.org/10.18632/oncotarget.9795>.
- [59] X. Wu, Q. Huang, R. Javed, J. Zhong, H. Gao, H. Liang, Effect of tobacco smoking on the epigenetic age of human respiratory organs, *Clin. Epigenet.* 11 (1) (2019) 183, <https://doi.org/10.1186/s13148-019-0777-z>.
- [60] H.K. Koo, J. Morrow, P. Kachroo, K. Tantisira, S.T. Weiss, C.P. Hersh, E. K. Silverman, D.L. DeMeo, Sex-specific associations with DNA methylation in lung tissue demonstrate smoking interactions, *Epigenetics* 16 (6) (2021) 692–703, <https://doi.org/10.1080/15592294.2020.1819662>.
- [61] K.W. Lee, Z. Pausova, Cigarette smoking and DNA methylation, *Front Genet.* 4 (2013) 132, <https://doi.org/10.3389/fgene.2013.00132>.
- [62] N.S. Shenker, P. Magne Ueland, S. Polidoro, K. van Veldhoven, F. Ricceri, R. Brown, J.M. Flanagan, P. Vineis, DNA methylation as a long-term biomarker of exposure to tobacco smoke, *Epidemiology* 24 (2013) 712–716, <https://doi.org/10.1097/EDE.0b013e31829d5cb3>.
- [63] R. Philibert, M. Dogan, A. Noel, S. Miller, B. Krukow, E. Papworth, J. Cowley, J. D. Long, S.R.H. Beach, D.W. Black, Dose response and prediction characteristics of a methylation sensitive digital pcr assay for cigarette consumption in adults, *Front. Genet.* 9 (2018) 137, <https://doi.org/10.3389/fgene.2018.00137>.
- [64] D. Wen, J. Shi, Y. Liu, W. He, W. Qu, C. Wang, H. Xing, Y. Cao, J. Li, L. Zha, DNA methylation analysis for smoking status prediction in the Chinese population based on the methylation-sensitive single-nucleotide primer extension method, *Forensic Sci. Int.* 339 (2022), 111412, <https://doi.org/10.1016/j.forsciint.2022.111412>.
- [65] N. Kondratyev, A. Golov, M. Alifimova, T. Lezheiko, V. Golimbet, Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation, *Clin. Epigenet.* 10 (130) (2018), <https://doi.org/10.1186/s13148-018-0565-1>.
- [66] P. de Knijff, From next generation sequencing to now generation sequencing in forensics, *Forensic Sci. Int. Genet.* 38 (2019) 175–180, <https://doi.org/10.1016/j.fsigen.2018.10.017>.
- [67] R. Vaisvila, V.K.C. Ponnaluri, Z. Sun, B.W. Langhorst, L. Saleh, S. Guan, N. Dai, M. A. Campbell, B.S. Sexton, K. Marks, M. Samaranyake, J.C. Samuelson, H. E. Church, E. Tamanaha, I.R. Corrêa Jr., S. Pradhan, E.T. Dimalanta, T.C. Evans Jr., L. Williams, T.B. Davis, Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA, in: *Genome Res.*, 31, 2021, pp. 1280–1289, <https://doi.org/10.1101/gr.266551.120>.
- [68] A. Heidegger, C. Xavier, H. Niederstätter, M. de la Puente, E. Pošpiech, A. Pisarek, M. Kayser, W. Branicki, W. Parson, Development and optimization of the VISAGE basic prototype tool for forensic age estimation, *Forensic Sci. Int. Genet.* 48 (2020), 102322, <https://doi.org/10.1016/j.fsigen.2020.102322>.
- [69] J.A. Sena, G. Galotto, N.P. Devitt, M.C. Connick, J.L. Jacobi, P.E. Umale, L. Vidali, C.J. Bell, Unique molecular identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis, *Sci. Rep.* 8 (1) (2018) 13121, <https://doi.org/10.1038/s41598-018-31064-7>.
- [70] K. Dawes, A. Andersen, R. Reimer, J.A. Mills, E. Hoffman, J.D. Long, S. Miller, R. Philibert, The relationship of smoking to cg05575921 methylation in blood and saliva DNA samples from several studies, *Sci. Rep.* 11 (1) (2021) 21627, <https://doi.org/10.1038/s41598-021-01088-7>.
- [71] R. Philibert, M. Dogan, S.R.H. Beach, J.A. Mills, J.D. Long, AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA, *Am. J. Med Genet. B Neuropsychiatr. Genet.* 183 (1) (2020) 51–60, <https://doi.org/10.1002/ajmg.b.32760>.
- [72] L. Grieshaber, S. Graw, M.J. Barnett, M.D. Thornquist, G.E. Goodman, C. Chen, D. C. Koestler, C.J. Marsit, J.A. Doherty, AHRR methylation in heavy smokers: associations with smoking, lung cancer risk, and lung cancer mortality, *BMC Cancer* 20 (1) (2020) 905, <https://doi.org/10.1186/s12885-020-07407-x>.
- [73] D.M. Tantoh, K.J. Lee, O.N. Nfor, Y.C. Liaw, C. Lin, H.W. Chu, P.H. Chen, S.Y. Hsu, W.H. Liu, C.C. Ho, C.C. Lung, M.F. Wu, Y.C. Liaw, T. Debnath, Y.P. Liaw, Methylation at cg05575921 of a smoking-related gene (AHRR) in non-smoking Taiwanese adults residing in areas with different PM2.5 concentrations, *Clin. Epigenet.* 11 (1) (2019) 69, <https://doi.org/10.1186/s13148-019-0662-9>.
- [74] D.M. Tantoh, M.C. Wu, C.C. Chuang, P.H. Chen, Y.S. Tyan, O.N. Nfor, W.Y. Lu, Y. P. Liaw, AHRR cg05575921 methylation in relation to smoking and PM2.5 exposure among Taiwanese men and women, *Clin. Epigenet.* 12 (1) (2020) 117, <https://doi.org/10.1186/s13148-020-00908-3>.
- [75] S.L. Park, Y.M. Patel, L.W.M. Loo, D.J. Mullen, I.A. Offringa, A. Maunakea, D. O. Stram, K. Siegmund, S.E. Murphy, M. Tiirikainen, L. Le Marchand, Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations, *Clin. Epigenet.* 10 (1) (2018) 110, <https://doi.org/10.1186/s13148-018-0543-7>.
- [76] R. Philibert, J.A. Mills, J.D. Long, S.E. Salisbury, A. Comellas, A. Gerke, K. Dawes, M. Vander Weg, E.A. Hoffman, The reversion of cg05575921 methylation in smoking cessation: a potential tool for incentivizing healthy aging, *Genes* 11 (12) (2020), <https://doi.org/10.3390/genes11121415>.
- [77] P.A. Dugue, C.H. Jung, J.E. Joo, X. Wang, E.M. Wong, E. Makalic, D.F. Schmidt, L. Baglietto, G. Severi, M.C. Southey, D.R. English, G.G. Giles, R.L. Milne, Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility, *Epigenetics* 15 (4) (2020) 358–368, <https://doi.org/10.1080/15592294.2019.1668739>.
- [78] V. Barcelona, Y. Huang, K. Brown, J. Liu, W. Zhao, M. Yu, S.L.R. Kardya, J.A. Smith, J.Y. Taylor, Y.V. Sun, Novel DNA methylation sites associated with cigarette smoking among African Americans, *Epigenetics* 14 (4) (2019) 383–391, <https://doi.org/10.1080/15592294.2019.1588683>.
- [79] H.R. Elliott, T. Tillin, W.L. McArdle, K. Ho, A. Duggirala, T.M. Frayling, G.D. Smith, A.D. Hughes, N. Chaturvedi, C. Relton, Differences in smoking associated DNA methylation patterns in South Asians and Europeans, *Clin. Epigenet.* 6 (4) (2014) 1–10, <https://doi.org/10.1186/1868-7083-6-4>.
- [80] C. You, S. Wu, S.C. Zheng, T. Zhu, H. Jing, K. Flagg, G. Wang, L. Jin, S. Wang, A. E. Teschendorff, A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes, *Nat. Commun.* 11 (1) (2020) 4779, <https://doi.org/10.1038/s41467-020-18618-y>.
- [81] M.M. Laqqan, M.M. Yassin, Cigarette heavy smoking alters DNA methylation patterns and gene transcription levels in humans spermatozoa, *Environ. Sci. Pollut. Res. Int.* 29 (18) (2022) 26835–26849, <https://doi.org/10.1007/s11356-021-17786-8>.
- [82] T.G. Jenkins, E.R. James, D.F. Alonso, J.R. Hoidal, P.J. Murphy, J.M. Hotaling, B. R. Cairns, D.T. Carrell, K.I. Aston, Cigarette smoking significantly alters sperm DNA methylation patterns, *Andrology* 5 (6) (2017) 1089–1099, <https://doi.org/10.1111/andr.12416>.
- [83] D.A. Gadd, A.J. Stevenson, R.F. Hillary, D.L. McCartney, N. Wrobel, S. McCafferty, L. Murphy, T.C. Russ, S.E. Harris, P. Redmond, A.M. Taylor, C. Smith, J. Rose, T. Millar, T.L. Spire-Jones, S.R. Cox, R.E. Marioni, Epigenetic predictors of lifestyle traits applied to the blood and brain, *Brain Commun.* 3 (2) (2021), fcab082, <https://doi.org/10.1093/braincomms/fcab082>.
- [84] Z. Yang, J. Yang, Y. Mao, M.D. Li, Investigation of the genetic effect of 56 tobacco-smoking susceptibility genes on DNA methylation and RNA expression in human brain, *Front. Psychiatry* 13 (2022), 924062, <https://doi.org/10.3389/fpsy.2022.924062>.

- [85] M.K. Lei, F.X. Gibbons, R.L. Simons, R.A. Philibert, S.R.H. Beach, The effect of tobacco smoking differs across indices of DNA methylation-based aging in an African American sample: DNA methylation-based indices of smoking capture these effects, *Genes* 11 (3) (2020), <https://doi.org/10.3390/genes11030311>.
- [86] H. Peng, W. Gao, W. Cao, J. Lv, C. Yu, T. Wu, S. Wang, Z. Pang, M. Yu, H. Wang, X. Wu, L. Li, Combined healthy lifestyle score and risk of epigenetic aging - a discordant monozygotic twin study, *Aging* 13 (10) (2021) 14039–14052, <https://doi.org/10.18632/aging.203022>.
- [87] Y. Yang, X. Gao, A.C. Just, E. Colicino, C. Wang, B.A. Coull, L. Hou, Y. Zheng, P. Vokonas, J. Schwartz, A.A. Baccarelli, Smoking-related DNA Methylation is associated with DNA methylation phenotypic age acceleration: the veterans affairs normative aging study, *Int. J. Environ. Res. Public Health* 16 (13) (2019), <https://doi.org/10.3390/ijerph16132356>.
- [88] C. Amador, Y. Zeng, M. Barber, R.M. Walker, A. Campbell, A.M. McIntosh, K. L. Evans, D.J. Porteous, C. Hayward, J.F. Wilson, P. Navarro, C.S. Haley, Genome-wide methylation data improves dissection of the effect of smoking on body mass index, *PLoS Genet.* 17 (9) (2021), e1009750, <https://doi.org/10.1371/journal.pgen.1009750>.
- [89] A. Andersen, R. Reimer, K. Dawes, A. Becker, N. Hutchens, S. Miller, M. Dogan, B. Hundley, A.M. J, D.L. J, R. Philibert, DNA methylation differentiates smoking from vaping and non-combustible tobacco use, *Epigenetics* 17 (2) (2022) 178–190, <https://doi.org/10.1080/15592294.2021.1890875>.
- [90] M. Bray, Y. Chang, T.B. Baker, D. Jorenby, R.M. Carney, L. Fox, G. Pham, F. Stoneking, N. Smock, C.I. Amos, L. Bierut, L.S. Chen, The promise of polygenic risk prediction in smoking cessation: evidence from two treatment trials, *Nicotine Tob. Res.* 24 (10) (2022) 1573–1580, <https://doi.org/10.1093/ntr/ntac043>.
- [91] X. Gao, H. Thomsen, Y. Zhang, L.P. Breitling, H. Brenner, The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes, *Clin. Epigenet.* 9 (2017) 87, <https://doi.org/10.1186/s13148-017-0387-6>.
- [92] Y. Xu, L. Cao, X. Zhao, Y. Yao, Q. Liu, B. Zhang, Y. Wang, Y. Mao, Y. Ma, J.Z. Ma, T. J. Payne, M.D. Li, L. Li, Prediction of smoking behavior from single nucleotide polymorphisms with machine learning approaches, *Front. Psychiatry* 11 (2020) 416, <https://doi.org/10.3389/fpsy.2020.00416>.
- [93] Z. Wang, A. Masoomi, Z. Xu, A. Boueiz, S. Lee, T. Zhao, R. Bowler, M. Cho, E. K. Silverman, C. Hersh, J. Dy, P.J. Castaldi, Improved prediction of smoking status via isoform-aware RNA-seq deep learning models, *PLoS Comput. Biol.* 17 (10) (2021), e1009433, <https://doi.org/10.1371/journal.pcbi.1009433>.
- [94] C. Díez López, D. Montiel González, A. Vidaki, M. Kayser, Prediction of smoking habits from class-imbalanced saliva microbiome data using data augmentation and machine learning, *Front. Microbiol.* 13 (2022), 886201, <https://doi.org/10.3389/fmicb.2022.886201>.